# Parte 2

# EL MODELO DE REGRESIÓN LOGÍSTICA

## CONTENIDO


1. Modelos de regresión con respuesta cualitativa..... 64

2. Fundamento del modelo de Regresión Logística ..... 65

3. El concepto de “odd” (o “riesgo”) ..... 66

4. Forma analítica del modelo de Regresión Logística..... 67

5. Interpretación de los parámetros del modelo de Regresión Logística

   5.1 Caso de una única variable de naturaleza cualitativa ..... 68
   5.2 Relación del MRL con el test de independencia ..... 69
   5.3 Caso de una única variable de naturaleza cuantitativa ..... 70
   5.4 Caso de dos variables explicativas. Interacciones..... 71

6. Estimación del Modelo de Regresión Logística..... 73

7. Inferencia en Modelos de Regresión Logística ..... 76

8. Test global de un ajuste. Análisis de las “deviances” ..... 78

9. Significación de un único parámetro: Método de Wald ..... 79

10. Estimación de una $LD_{50}$ de un tóxico ..... 80

11. Predicción de la afectación del ganglio centinela..... 83

12. Factores pronóstico en la enfermedad hemolítica neonatal ..... 88

13. Modelo de Regresión Multinomial..... 90

## 1. MODELOS DE REGRESIÓN CON RESPUESTA CUALITATIVA

El Modelo de Regresión Clásico asume que el valor $y_j$ observado en cada individuo para la variable dependiente **Y**, es una variable aleatoria normal, de varianza $\sigma^2(Y_j)$ constante desconocida, y cuyo valor medio es una función de los valores $x_{ij}$ constatados en ese individuo para las variables explicativas $X_i$

$$\mathbf{E(Y_j) = f(x_{1j}, \ldots, x_{Ij})}$$

En ciertas situaciones, sin embargo, la variable respuesta **Y** de interés, puede ser de naturaleza cualitativa, sea de tipo binario (con sólo dos alternativas posibles), sea con un conjunto de "valores" posibles cualitativamente diferentes.

Por ejemplo, en un estudio sobre la efectividad en niños de una nueva vacuna contra cierta enfermedad, el "valor" observado en cada individuo sería si éste sufre o no finalmente esta enfermedad.

En este caso la variable respuesta $y_j$ observada en cada niño será una variable binaria con sólo dos alternativas: 0 (el niño no padece la enfermedad) ó 1 (el niño sí padece la enfermedad)

Como otro ejemplo, en un estudio sobre la efectividad de diversas alternativas para el tratamiento de cierta enfermedad, la variable respuesta en cada individuo podría tener cuatro alternativas, según que éste: empeorase, se quedara igual, mejorase o curara definitivamente

En esta Parte 2 del curso se estudia con detalle el **Modelo de Regresión Logística (MRL)** que permite analizar situaciones en la que la respuesta es de naturaleza binaria.

Finalmente se expone, de manera sucinta, cómo es posible generalizar el MRL para abordar situaciones en las que la variable respuesta tiene más de dos alternativas cualitativamente diferentes.

## 2. FUNDAMENTO DEL MODELO DE REGRESIÓN LOGÍSTICA

El Modelo de Regresión Logística se utiliza en situaciones en las que la variable dependiente de interés **Y** no es de naturaleza continua, sino que simplemente indica si se ha producido o no un determinado suceso (por ejemplo, si un infarto ha originado o no el fallecimiento del infartado)

La variable respuesta asociada a unos determinados valores $x_{ij}$ de las variables explicativas sólo tiene por tanto dos "valores" posibles, que se pueden codificar como:

- Y = 1 si se he producido el suceso estudiado (en el ejemplo: si la persona infartada fallece en un determinado plazo)

- Y = 0 si no se produce el suceso (en el ejemplo: si la persona infartada no fallece en el plazo considerado)

La distribución de este tipo de variable aleatoria (**Distribución de Bernouilli**) viene caracterizada por un único parámetro **p**, que es la probabilidad de que se produzca el suceso en cuestión.

Se demuestra fácilmente que la media E(**Y**) de **Y** es precisamente **p**

Por lo tanto, en estas situaciones el modelo general de regresión es:

$$E(Y/X_1=x_{1j},\ldots,X_l=x_{lj}) = \mathbf{p}/\underline{(X_1=x_{1j},\ldots,X_l=x_{lj})} = f(x_{1j}\ldots x_{lj}) \qquad (1)$$

**El efecto de las variables explicativas en el MRL consiste en modificar la probabilidad de que se produzca un determinado suceso**

## 3. EL CONCEPTO DE “ODD” (O “RIESGO”)

En el modelo (1) se presenta el problema de que la función $f(x_{1j}\ldots x_{lj})$ deberá estar siempre comprendida entre 0 y 1, al tratarse de una probabilidad.

Para obviar este problema se utiliza el concepto de Odd[1].

El “odd” (OD) asociado a un determinado suceso de probabilidad P, se define como el cociente P/(1-P).

$$\text{odd} = \frac{P}{1-P}$$

Así, por ejemplo, a un suceso de probabilidad P=0.8 le correspondería un OD igual a 0.8/(1-0.8)=4, porque es 4 veces más probable que se produzca que que no se produzca.

En la literatura médica se utiliza frecuentemente el término “riesgo” para referirse a los “odds”, de forma que, con esta nomenclatura, riesgo de una determinada patología en una población es igual a prevalencia/(1-prevalencia) donde la prevalencia viene expresada en tanto por uno (y no en porcentaje, como se hace habitualmente).

Los OD pueden variar entre 0 (cuando P=0) e infinito (cuando P=1) y, por lo tanto, su logaritmo (se trabaja habitualmente con logaritmos neperianos) puede tomar cualquier valor entre $-\infty$ (que es el logaritmo de 0) y $+\infty$ (que es el logaritmo de $+\infty$).

[1] Como no existe un consenso generalizado sobre la traducción al castellano del “odd” (en algún texto se utiliza para ello el término “momio”) y dado que dicho término se utiliza ampliamente en la literatura técnica, hemos decidido mantenerlo sin traducir.

## 4. FORMA ANALÍTICA DEL MODELO DE REGRESIÓN LOGÍSTICA

En los Modelos de Regresión Logística (MRL) la ecuación de regresión modeliza la relación entre el logaritmo neperiano de los OD y las variables explicativas:

$$\ln\left(\frac{p}{1-p}\right) = f(x_{1j}\ldots x_{Ij}) \qquad (2)$$

no habiendo ahora problemas con que "f" tome cualquier valor entre -∞ y +∞

Como en los modelos clásicos de regresión lineal, la función "f" puede escribirse en la forma general:

$$f(x_{1j},\ldots,x_{Ij}) = \sum_{k=1}^{k=p} \beta_k z_{kj} \qquad (3)$$

Donde las $z_k$, a los que denominaremos "regresores" son k funciones de las $\mathbf{x}_i$,

$$z_k = g_k(\mathbf{x}) = g_k(x_1,\ldots,x_I) \qquad k = 1\ldots p$$

que incluirán, en general, las $x_j$ iniciales y transformaciones como sus cuadrados y productos, posibles variables binarias (o "dummy") y sus productos por otras (si hay factores cualitativos) y una primera variable de valor constante e igual a 1 que multiplicaría a un parámetro $\beta_0$.

De las ecuaciones (2) y (3) se deduce que en el MRL la probabilidad de que se produzca un determinado suceso se relaciona con los valores de las $X_i$ mediante la expresión:

$$(\mathbf{p}/X_1=x_{1j},\ldots,X_I=x_{Ij}) = \frac{e^{\sum_{k=1}^{k=p}\beta_k z_{kj}}}{1+e^{\sum_{k=1}^{k=p}\beta_k z_{kj}}} \qquad (4)$$

# 5. INTERPRETACIÓN DE LOS PARÁMETROS EN EL MODELO DE REGRESIÓN LOGÍSTICA

## 5.1 Caso de una única variable de naturaleza cualitativa

Sea **p** la probabilidad de fallecer en enfermos que sufren determinada enfermedad. Se desea estudiar como influye sobre **p** el tratamiento aplicado, comparándose tres tratamientos alternativos A, B y C.

Para introducir en la ecuación (4) el efecto del tratamiento utilizado sobre **p**, se definen, como en los modelos de regresión clásicos, dos variables "dummy" de la siguiente forma:

$X_1$ vale 1 si se aplica el tratamiento B y vale 0 en caso contrario

$X_2$ vale 1 si se aplica el tratamiento C y vale 0 en caso contrario

El modelo de regresión logística es:

$$\log\frac{P}{1-P} = \beta_o + \beta_1 x_1 + \beta_2 x_2 \quad \Rightarrow \quad \frac{P}{1\text{-}P} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2} = e^{\beta_0} e^{\beta_1 x_1} e^{\beta_2 x_2}$$

y definiendo el riesgo de fallecimiento por el odd P/(1-P) se tiene:

Riesgo de fallecimiento con Tratamiento A ($x_1$=0, $x_2$=0): $e^{\beta_0}$
Riesgo de fallecimiento con Tratamiento B ($x_1$=1, $x_2$=0): $e^{\beta_0} e^{\beta_1}$
Riesgo de fallecimiento con Tratamiento C ($x_1$=0, $x_2$=1): $e^{\beta_0} e^{\beta_2}$

Por lo tanto:

$e^{\beta_1}$ : es el valor por el que se multiplica el riesgo si se utiliza el tratamiento B en vez del tratamiento A.
$e^{\beta_2}$ : es el valor por el que se multiplica el riesgo si se utiliza el tratamiento C en vez del tratamiento A.

Así, por ejemplo, si $\beta_1$ sale positivo $e^{\beta_1}$ es mayor que 1 y el riesgo aumenta utilizando B en vez de A, mientras que si $\beta_1$ sale negativo $e^{\beta_1}$ es menor que 1 y el riesgo utilizando B es menor que con A

Si $\beta_1$ y $\beta_2$ fueran iguales a 0, como $e^0 = 1$, los tres riesgos resultarían iguales y el tratamiento aplicado no modificaría la probabilidad de fallecimiento.

## 5.2 Relación del MRL con el test de independencia en tablas de frecuencias

Los datos para estimar el modelo del ejemplo anterior tendrían el aspecto de una tabla de frecuencias clásica con 2 columnas:

| | Fallecidos | No Fallecidos |
|---|---|---|
| Tratamiento A | $x_{11}$ | $x_{12}$ |
| Tratamiento B | $x_{21}$ | $x_{22}$ |
| Tratamiento C | $x_{31}$ | $x_{32}$ |

El clásico test Gi-dos de independencia entre filas y columnas contrasta la hipótesis de que la probabilidad de fallecimiento no depende del tratamiento, o sea que es la misma para los tres tratamientos,

Este test es equivalente al que se hace en el MRL visto en el ejemplo anterior para contrastar la hipótesis $\beta_1 = \beta_2 = 0$. (La forma de contrastar hipótesis en el MRL se expone más adelante)

Un ventaja del MRL frente al test clásico es que, en el caso de rechazar la hipótesis de independencia permite analizar directamente (a partir de la significación individual de $\beta_1$, $\beta_2$ y $\beta_1$-$\beta_2$) entre qué tratamientos existen diferencias significativas en las probabilidades de supervivencia.

Adicionalmente, y ésta es la ventaja más importante, el MRL permite analizar tablas de frecuencias múltiples, por ejemplo la que aparecería si se dispusiera de dos tablas como la anterior: una para varones y otra para mujeres. En estos casos el MRL permite investigar además (como se verá más adelante) la posible existencia de interacciones (como, por ejemplo, que el efecto favorable de un determinado tratamiento sea más marcado en varones que en mujeres)

El MRL permite incorporar también el estudio de variables de naturaleza cuantitativa, como por ejemplo la edad del paciente, sin necesidad de discretizarla en un conjunto de grupos (por ejemplo: <30  30-60  >60)

Cuando la variable respuesta tiene más de 2 alternativas (más de 2 columnas en la tabla de frecuencias) es posible abordar el estudio mediante un Modelo de Regresión Logística Multinomial (ver al final de esta Parte 2)

## 5.3 Caso de una única variable de naturaleza cuantitativa

Sea **p** la probabilidad de fallecer a causa de un determinado tumor cancerígeno en mama. Se desea estudiar cómo influye sobre **p** el tamaño X del tumor (medido en cms en el momento de la masectomía)

En este caso X puede introducirse directamente en la ecuación del MRL.

$$\log\frac{P}{1-P} = \beta_o + \beta_1 x \quad \Rightarrow \quad \frac{P}{1\text{-}P} = e^{\beta_0+\beta_1 x}$$

De acuerdo con la anterior expresión, el aumento de 1 cm del tamaño del tumor hace que el riesgo pase de:

$$e^{\beta_0+\beta_1 x} \Rightarrow e^{\beta_0+\beta_1(x+1)} = e^{\beta_0+\beta_1 x}e^{\beta_1}.$$

Por lo tanto $e^{\beta_1}$ cuantifica por cuanto se multiplica el riesgo por cada cm de aumento en el tamaño del tumor, o sea es el ratio (o.r.) entre los OD después y antes de aumentar una unidad $x_1$.

Es posible que el efecto del tamaño del tumor sobre el logaritmo del riesgo sea no lineal, por ejemplo que el paso de 3 a 5 cm tengo un efecto más marcado sobre ln(o.r.) que el paso de 1 a 3 cm. Si se desea que el modelo contemple esta posibilidad podría postularse, como en los modelos de regresión clásicos, el siguiente modelo alternativo:

$$\log\frac{P}{1-P} = \beta_o + \beta_1 x + \beta_2 x^2$$

Un valor de $\beta_2$ positivo implicaría que el aumento del tamaño del tumor tiene un efecto cada vez mayor sobre el riesgo de mortalidad, mientras que un valor de $\beta_2$ negativo indicaría lo contrario.

## 5.4 Caso de dos variables explicativas. Interacciones

Sea **p** la probabilidad de fallecer a causa de un determinado tumor cancerígeno en mama. Se desea estudiar como influye sobre **p** el tamaño $x_1$ del tumor (medido en cm) y la aplicación o no de un determinado tratamiento A (por ejemplo, aplicar quimioterapia antes de la intervención quirúrgica).

Para modelizar el posible efecto del tratamiento se define una única variable binaria $X_A$ (dado que sólo se contemplan dos alternativas) de forma que $X_A$ vale 1 si se aplica el tratamiento A y vale 0 si no se aplica.

En el siguiente MRL sencillo:

$$\log\frac{P}{1-P} = \beta_o + \beta_1 x_1 + \beta_A x_A \quad \Rightarrow \quad \frac{P}{1\text{-}P} = e^{\beta_0+\beta_1 x_1+\beta_A x_A} = e^{\beta_0} e^{\beta_1 x_1} e^{\beta_A x_A}$$

el riesgo, en función del tamaño $x_1$ del tumor, tiene las siguientes expresiones, según se aplique o no el tratamiento A:

- Riesgo si no se aplica el tratamiento a ($x_A$ = 0) = $e^{\beta_0} e^{\beta_1 x_1} e^{\beta_A x0} = e^{\beta_0} e^{\beta_1 x_1}$
- Riesgo si se aplica el tratamiento a ($x_A$ = 1) = $e^{\beta_0} e^{\beta_1 x_1} e^{\beta_A x1} = e^{\beta_0} e^{\beta_1 x_1} e^{\beta_A}$

Por lo tanto el efecto de aplicar el tratamiento A será, sea cual sea el tamaño $x_1$ del tumor, multiplicar el riesgo por $e^{\beta_A}$ (que, por ejemplo, si $\beta_A$ fuera negativo sería un número menor que 1 y, por tanto, el riesgo disminuiría).

Podría suceder, sin embargo, que el tratamiento A sólo fuera favorable si los tumores son grandes, o sea que el efecto de dicho tratamiento dependiera del tamaño $x_1$ del tumor.

Se dice en estos casos que hay una **interacción** entre los efectos de ambas variables sobre el riesgo.

**Consideración de interacciones en el MRL (continuación)**

La posibilidad de existencia de interacciones puede contemplarse en el MRL, introduciendo, como se hace en los modelos clásicos de regresión, una nueva variable definida como el producto de las dos que pueden interaccionar.

El nuevo MRL sería:

$$\log \frac{P}{1-P} = \beta_o + \beta_1 x_1 + \beta_A x_A + \beta_{1A} x_1 x_A$$

$$\downarrow$$

$$\frac{P}{1\text{-}P} = e^{\beta_0 + \beta_1 x_1 + x_A(\beta_A + \beta_{1A} x_1)} = e^{\beta_0} e^{\beta_1 x_1} e^{x_A(\beta_A + \beta_{1A} x_1)}$$

De acuerdo con este modelo el efecto sobre el riesgo de aplicar el tratamiento A consistiría en multiplicarlo por $e^{\beta_A + \beta_{1A} x_1}$, con lo que, según los valores resultantes de $\beta_A$ y $\beta_{1A}$, dicha aplicación podría, por ejemplo, ser perjudicial para tumores pequeños y beneficiosa para tumores grandes.

También es posible considerar interacciones en un modelo más complejo que asumiera que el efecto del tamaño del tumor sobre el logaritmo del riesgo no es lineal, debiendo entonces introducirse, además de la variable $x_1^2$, dos variables $x_1 x_A$ y $x_1^2 x_A$ para modelizar la posible interacción.

## 6. ESTIMACIÓN DEL MODELO DE REGRESIÓN LOGÍSTICA

La estimación de los parámetros $\beta_k$ en el MRL (4) no puede llevarse a cabo por el método ordinario de mínimos cuadrados por dos motivos:

- la ecuación del modelo (4) no es una función lineal en los parámetros, por lo que la obtención de los estimadores $b_k$ que minimizan la suma de cuadrados de las diferencias entre los valores observados $Y_j$ y su valores medios $H(b_1,\ldots,b_k)$ no conduce a un sistema de ecuaciones lineales fácilmente resoluble

- La distribución de las $Y_j$ no es normal, sino de Bernouilli, con lo que la eventual minimización de dicha suma de cuadrados no conduciría a estimadores estadísticamente óptimos

En general, el software estadístico estima los MRL aplicando directamente el **método de Máxima Verosimilitud**

El fundamento de dicho método radica en tomar como estimadores de los parámetros desconocidos aquellos valores que hacen máxima la **Verosimilitud** de la muestra, o sea la probabilidad de haber observado precisamente los datos que se han obtenido.

La obtención de los valores de las $\beta_k$ maximiza la verosimilitud V, se simplifica si se trabaja con el logaritmo de V, puesto que, al ser el logaritmo una función monótona creciente, los valores $\beta_i^*$ que maximicen ln(V), maximizarán también V, y los cálculos resultan mucho más sencillos al transformarse los productos en sumas.

Siendo V* el valor máximo de V obtenido en la estimación, a -2ln(V*) se le denomina la ***deviance residual*** **($D_{res}$)** del ajuste, jugando en estos modelos, como se verá más adelante, un papel similar al que juega la SCresidual en los ajustes de regresión ordinarios[2].

Como subproducto de este proceso de optimización se obtiene también una estimación de las varianzas de los estimadores obtenidos, que permite la realización de inferencias respecto a los mismos.

---

[2] Obsérvese que, en efecto y al igual de lo que sucede con la $SC_{residual}$ en la regresión clásica, cuanto mejor es el ajuste (y más cerca de 1 está por tanto la verosimilitud V*) más cerca de 0 está la "deviance" -2ln(V*)

**Nota técnica sobre el método de Máxima Verosimilitud**

(Esta sección es de carácter más técnico, exigiendo ciertos conocimientos previos estadísticos para su comprensión, siendo ésta no indispensable para seguir el resto del tema)

Ejemplo elemental aclaratorio

En N observaciones se ha constatado que un determinado suceso se ha presentado x veces. ¿Cuál será el estimador máximo verosímil p* de la probabilidad p de dicho suceso?

Como el suceso se ha presentado x veces y no se ha presentado las N-x veces restantes, la probabilidad de que esto suceda (en el orden concreto en el que han sucedido las presencias y ausencias) será:

$$V = \Pr ob(X = x) = p^x (1-p)^{N-x}$$

La obtención del valor p* de p que maximiza la expresión anterior se simplifica si se trabaja con el logaritmo neperiano L de V, puesto que los cálculos resultan mucho más sencillos L = ln(V) = xln(p) + (N-x)ln(1-p)

Igualando a cero la derivada respecto a p para obtener el máximo se obtiene inmediatamente

$$\frac{dL}{dp} = \frac{x}{p} - \frac{N-x}{1-p} = 0 \quad \text{(y operando)} \rightarrow p^* = \frac{x}{N}$$

resultando, como era de prever, que el estimador máximo verosímil p* de p es precisamente la frecuencia relativa observada para el suceso.

Propiedades de los estimadores máximos verosímiles

Los estimadores obtenidos por el Método de Máxima Verosimilitud gozan en general de buenas propiedades estadísticas, especialmente si los tamaños muestrales son grandes.

Se demuestra, en efecto, que el estimador máximo verosímil $\vec{b}$ de un determinado vector de k parámetros $\vec{\beta}$ se distribuye asintóticamente de forma normal, siendo asintóticamente insesgado y teniendo una matriz de varianzas-covarianzas "menor o igual"[3] que la de cualquier otro estimador que sea asintóticamente insesgado.

Se demuestra también que asintóticamente la matriz var-cov de $\vec{b}$ coincide con la cota de Cramer-Rao multivariante, que es la inversa de la matriz de información, definida esta última como la matriz kxk cuyo término general (i,j) es la esperanza matemática de $-\frac{\partial^2 L}{\partial \beta_i \partial \beta_j}$

[3] Decimos que una matriz A es "menor" que otra B, si A = B + C siendo C una matriz definida o semidefinida positiva.

**Nota técnica sobre el método de Máxima Verosimilitud (continuación)**

Estimación máximo verosímil del MRL

Para estimar el MRL definido por (4)

$$(\mathbf{p}/X_1=x_{1j},\ldots,X_l=x_{lj}) = \frac{e^{\sum_{k=1}^{k=p}\beta_k z_{kj}}}{1+e^{\sum_{k=1}^{k=p}\beta_k z_{kj}}}$$

se dispone en general para cada uno de los N "individuos" de la muestra (o sea para j=1,…,N) de los valores de las variables explicativas $\mathbf{x_j} = \{x_{1j},\ldots,x_{lj}\}$ y del valor de $Y_j$, donde $Y_j = 1$ si en el individuo j se ha producido el suceso considerado o $Y_j = 0$ si no se ha producido.

La verosimilitud asociada a una observación $Y_j$ será igual a $\mathbf{p_j}$ si $Y_j=1$ o a $(1-\mathbf{p_j})$ si $Y_j=0$, pudiendo ponerse de la forma general $p_j^{Y_j}(1-p_j)^{(1-Y_j)}$, donde $\mathbf{p_j}$ está relacionado con las $\beta_k$ mediante (4), dado que los valores de las $z_{kj}$ en la observación son conocidos.

El logaritmo de la verosimilitud conjunta de los N valores observados $Y_j$ vendrá dada por la ecuación:

$$\ln V = \ln\left(\prod_{j=1}^{j=N}\left(p_j^{Y_j}(1-p_j)^{(1-Y_j)}\right)\right) = \sum_{j=1}^{j=n}\left[Y_j \ln p_j + (1-Y_j)\ln(1-p_j)\right]$$

que, sustituyendo las $\mathbf{p_j}$ por la expresión (4) y las $Y_j$ y $x_{ij}$ por los valores observados en cada individuo, acaba resultando una función $\ln V = H(\beta_1,\ldots,\beta_k)$ de los k parámetros desconocidos del modelo.

Los estimadores máximo-verosímiles de los parámetros $\beta_k$ se obtienen calculando, mediante técnicas de optimización numérica, los valores de dichos parámetros que maximizan la expresión anterior

Como subproducto de este proceso de optimización se obtiene también una estimación de la matriz de información, evaluada sustituyendo en ella los parámetros por sus estimadores, cuya inversa permite estimar la matriz de varianzas-covarianzas de dichos estimadores.

## 7. INFERENCIA EN MODELOS DE REGRESIÓN LOGÍSTICA

### Test de la razón de verosimilitudes generalizada

Una vez estimados los parámetros del modelo, la realización de inferencias respecto a los mismos puede llevarse a cabo utilizando el **test de la razón de verosimilitudes generalizada (TRVG)**

El TRVG es un procedimiento de inferencia de carácter general, para estudiar si es admisible la hipótesis de que los parámetros de un modelo cumplen o no un determinado conjunto de restricciones.

Ejemplo 1: en el MRL considerado al exponer el estudio de interacciones, se desea estudiar si es admisible la hipótesis $\beta_{1A} = 0$, que implicaría que no habría interacción entre los efectos sobre el riesgo de fallecimiento del tratamiento A y del tamaño del tumor.

Ejemplo 2: en el MRL considerado en el estudio del efecto de 3 posibles tratamientos, se desea investigar si es admisible la hipótesis $\beta_1 = 0$ y $\beta_2 = 0$ que implicaría que no hay diferencias entre los tres tratamientos

Siendo V la verosimilitud de la muestra (que depende del valor que tengan los parámetros del modelo), el TRVG se basa en el valor de un cociente **rv**, cuyo denominador es el máximo valor posible de dicha verosimilitud (que se obtiene sustituyendo los parámetros por sus estimadores máximo verosímiles) y cuyo numerador es el máximo valor posible de la verosimilitud cuando se impone a los parámetros el respetar la restricción a contrastar

$$\mathbf{rv} = \frac{\max(\text{V / modelo restringido})}{\max(\text{V / modelo no restringido})}$$

El ratio de verosimilitudes **rv** está siempre comprendido entre 0 y 1, siendo tanto más pequeño cuanto menos verosímil, a la luz de los datos disponibles, la restricción impuesta a los parámetros.

Por lo tanto, una determinada hipótesis (por ejemplo, $\beta_{1A} = 0$) se rechazará si el correspondiente **rv** resulta “demasiado pequeño”.

¿Y qué debemos considerar como ”demasiado pequeño”?

**Test de la razón de verosimilitudes generalizada (continuación)**

Se demuestra en condiciones muy generales que, cuando los tamaños muestrales crecen el estadístico -2ln(**rv**) tiende a distribuirse como una variable $\chi^2_r$ (siendo r el número de restricciones impuestas en ,los parámetros) si es cierta la hipótesis nula implicada por las restricciones consideradas, y tiende a ser mayor que una $\chi^2_r$ si dicha hipótesis es falsa.

Por tanto, denominando $\mathbf{D_{res}}$ a - 2ln(V) a la "**deviance residual"** del ajuste de un modelo

$$-2\ln(\mathbf{rv}) = D_{\text{res/odelo restringido}} - D_{\text{res/odelo no restringido}} \sim \chi^2_r$$

si son ciertas las restricciones impuestas a los parámetros

Por tanto la hipótesis contemplada se rechazará si el valor de -2ln(**rv**), o sea el <u>incremento en la "deviance residual" al imponer las r restricciones</u> a los parámetros, sale "demasiado grande" para ser una $\chi^2_r$ (o sea, si excede el percentil (1-$\alpha$) de dicha distribución, donde $\alpha$ es el riesgo de 1ª especie que se esté dispuesto a asumir.

En general cualquier software estadístico al estimar un MRL proporciona entre sus resultados la **deviance residual** del modelo.

Para obtener dicha **deviance residual** cuando se imponen restricciones a los parámetros, se estimará un modelo sin restricciones pero cuyo planteamiento lleve implícito el cumplimiento de la, o las, restricciones a considerar.

Por ejemplo, en el modelo discutido en el ejemplo expuesto sobre interacciones, plantear el modelo $\log\frac{P}{1-P} = \beta_o + \beta_1 x_1 + \beta_A x_A$ sin restricciones, pero sin incluir la variable $x_1x_A$, es equivalente a plantear el modelo completo $\log\frac{P}{1-P} = \beta_o + \beta_1 x_1 + \beta_A x_A + \beta_{1A} x_1 x_A$ pero incluyendo la restricción $\beta_{1A} = 0$.

## 8. TEST GLOBAL DE UN AJUSTE: ANÁLISIS DE LAS DEVIANCES

En el Modelo de Regresión Logística $\ln\frac{p}{1-p} = \beta_0 + \beta_1 z_1 + \ldots + \beta_I z_I$ la hipótesis $\beta_1 = \ldots = \beta_I = 0$ implicaría que ninguna de las variables explicativas contempladas influye sobre p, y el modelo tendría un único parámetro $\beta_0$

Si esta hipótesis es cierta, la estimación máximo-verosímil p* de p es simplemente la frecuencia observada del suceso en las N observaciones, y el valor correspondiente de la "deviance" -2ln(V*) es la **deviance total ($D_{tot}$)** de los datos (que depende sólo del número de observaciones y del número de ocurrencias del suceso considerado en la muestra). $D_{tot}$ tiene N-1 grados de libertad (número de datos – número de parámetros del modelo)

Una vez ajustado el modelo, sin imponer restricciones a las $\beta_i$, el valor obtenido de -2ln(V*) es la **deviance residual ($D_{res}$)** del ajuste. $D_{res}$ tiene N – (I+1) grados de libertad (número de datos – número de parámetros del modelo).

La diferencia **$D_{mod} = D_{tot} - D_{res}$** es la **deviance explicada por el modelo**, y tiene I grados de libertad, que es el número de variables explicativas.

Una vez ajustado un modelo, se puede obtener el siguiente cuadro resumen de las deviances, cuya función es muy similar a la del cuadro resumen del Anova en un modelo de regresión ordinario:

**Análisis de Desviación**

| *Fuente* | *Desviación* | *Gl* | *Valor-P* |
|---|---|---|---|
| Modelo | $D_{mod}$ | I | p-$val_{mod}$ |
| Residuo | $D_{res}$ | N-1-I | p-$val_{res}$ |
| Total (corr.) | $D_{tot}$ | N-1 | |

(Nota: Un modelo que tuviese tantos parámetros con observaciones existen[4] se ajustaría perfectamente a los datos y tendría una $D_{res}$ nula)

p-$val_{mod}$ permite estudiar si el modelo es globalmente significativo o sea, si no es admisible la hipótesis $\beta_1 = \ldots = \beta_I = 0$

p-$val_{res}$ permite estudiar si la variabilidad residual no explicada por el modelo es estadísticamente significativa (comparada con la de un modelo perfecto, con tantos parámetros como observaciones, que tendría deviance = 0)

[4] A estos efectos, cuando los datos se dan agrupados, proporcionando la frecuencia relativa observada para cada conjunto de valores posibles de las $x_i$, cada una de estas frecuencias relativas se considera como una única observación

## 9. SIGNIFICACIÓN DE UN ÚNICO PARÁMETRO $\beta_K$

### Método de Wald

Cuando el interés radica únicamente en estudiar la significación de un determinado parámetro $\beta_k$, o sea, en ver si es admisible la hipótesis $\beta_k = 0$, además del método del TRVG, es más sencillo utilizar el **método de Wald**.

El test de Wald consiste en dividir el valor estimado $b_k$ por su desviación típica estimada $s(b_k)$. La hipótesis $\beta_k = 0$ se rechaza si el valor obtenido es mayor que el percentil 1-$\alpha$ para una variable normal tipificada[5], donde $\alpha$ es el riesgo de 1ª especie que se esté dispuesto a asumir.

### Intervalo de confianza para el valor previsto de p cuando $\vec{z} = \vec{z}_0$

El MRL puede escribirse en forma vectorial como $\ln\frac{p}{1-p} = \sum_k \beta_k z_k = \vec{z}'\vec{\beta}$

La estimación máximo-verosímil del vector de parámetros $\vec{\beta}$ da como resultado el vector $\vec{b}$ de estimadores, así como su matriz de varianzas-covarianzas estimada $\mathbf{V_b}$

Dado un determinado valor $\vec{z}_0$ para el vector de los regresores, la estimación del valor previsto de $\ln\frac{p}{1-p}$ será $\vec{z}'_0\vec{b}$ , y un intervalo de confianza aproximado para dicho valor previsto será $\vec{z}'_0\vec{b} \pm t_\alpha \vec{z}'_0 V_b \vec{z}_0$ (donde $t_a$ es el percentil de la t de Student para el valor deseado del riesgo $\alpha$ de 1ª especie)

A partir del intervalo de confianza [$lo_1$ $lo_2$] obtenido para lo = $\ln\frac{p}{1-p}$, se obtiene el correspondiente intervalo de confianza para p: $\left[\frac{e^{lo_1}}{1+e^{lo_1}}; \frac{e^{lo_2}}{1+e^{lo_2}}\right]$

[5] Teóricamente sería más correcto comparar con el percentil de una t de Student con N-K grados de libertad, pero dado el carácter aproximado del test y que N-K es generalmente grande se suele trabajar con la normal tipificada

## 10. ESTIMACIÓN DE LA $LD_{50}$ DE UN TÓXICO

El archivo *escarabajos* contiene un conjunto de datos de Bliss (1935) que muestra los resultados de experimentos en los cuales se expusieron escarabajos a diferentes concentraciones de bisulfuro de carbono. El archivo de datos muestra la dosis, el número de escarabajos expuestos y el número de escarabajos muertos.

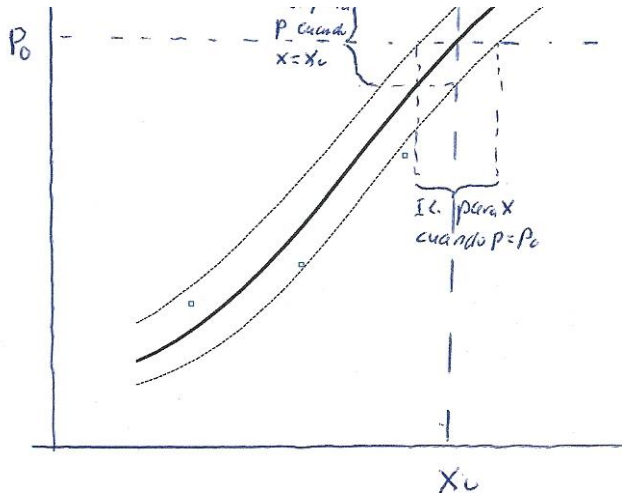
| Dosis | Expuestos | Muertos |
|---|---|---|
| 1.691 | 59 | 6 |
| 1.724 | 60 | 13 |
| 1.755 | 62 | 18 |
| 1.784 | 56 | 28 |
| 1.811 | 63 | 52 |
| 1.837 | 59 | 53 |
| 1.861 | 62 | 61 |
| 1.884 | 60 | 60 |

El objetivo del estudio es determinar la $LD_{50}$ del producto para los escarabajos (dosis del mismo que ocasiona la muerte del 50% de los expuestos), obteniendo un intervalo de confianza (con $\alpha$=5%) para el mismo.

Nota: Intervalo de confianza inverso

En un MRL con una sola variable explicativa x, se ha visto cómo es posible obtener un intervalo de confianza para el valor previsto de p cuando $x = x_0$.

A partir de dichos intervalos de confianza es posible, tal como se ve en la figura, resolver el problema inverso de hallar un intervalo de confianza para el valor $x_0$ de x que produce un valor determinado $p_0$ para p



## Estimación de una $LD_{50}$ (continuación)

<u>Modelo 1</u>: $\ln\frac{p}{1-p} = \beta_0 + \beta_1 x$ (siendo p la probabilidad de muerte cuando la dosis empleada es x)

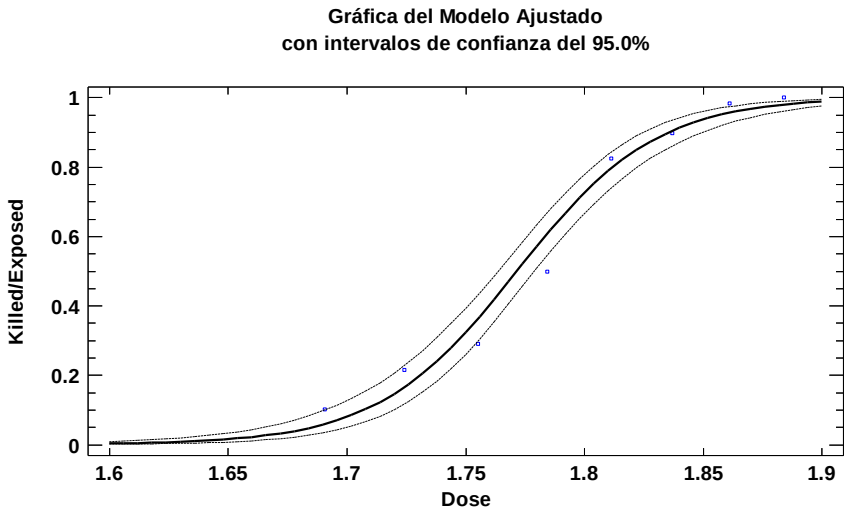
**Modelo Estimado de Regresión (Máxima Verosimilitud)**

| | | *Error* | *Razón de Momios* |
|---|---|---|---|
| *Parámetro* | *Estimado* | *Estándar* | *Estimada* |
| CONSTANTE | -60.7175 | 5.18048 | |
| Dose | 34.2703 | 2.91201 | 7.64563E14 |

**Análisis de Desviación**

| *Fuente* | *Desviación* | *Gl* | *Valor-P* |
|---|---|---|---|
| Modelo | 272.97 | 1 | 0.0000 |
| Residuo | 11.2322 | 6 | 0.0815 |
| Total (corr.) | 284.202 | 7 | |

Porcentaje de desviación explicado por el modelo = 96.0478




Aunque el efecto de la dosis es muy significativo, en la figura se observa que, en el intervalo analizado de dosis, los residuos son positivos para valores bajos de ésta, negativos para valores intermedios y nuevamente positivos para valores altos de dosis.

Tal como se vió al estudiar el Modelo de Regresión Clásico, esta estructura de los residuos hace sospechar que la relación entre $\ln\frac{p}{1-p}$ y x pueda ser no lineal.

## Estimación de una $LD_{50}$ (continuación)

Modelo 2: $$\ln\frac{p}{1-p} = \beta_0 + \beta_1 x + \beta_2 x^2$$

**Modelo Estimado de Regresión (Máxima Verosimilitud)**

| | | *Error* | *Razón de Momios* |
|---|---|---|---|
| *Parámetro* | *Estimado* | *Estándar* | *Estimada* |
| CONSTANTE | 431.106 | 180.632 | |
| Dose | -520.615 | 204.497 | 7.93752E-227 |
| Dose^2 | 156.412 | 57.8558 | 8.48545E67 |

**Análisis de Desviación**

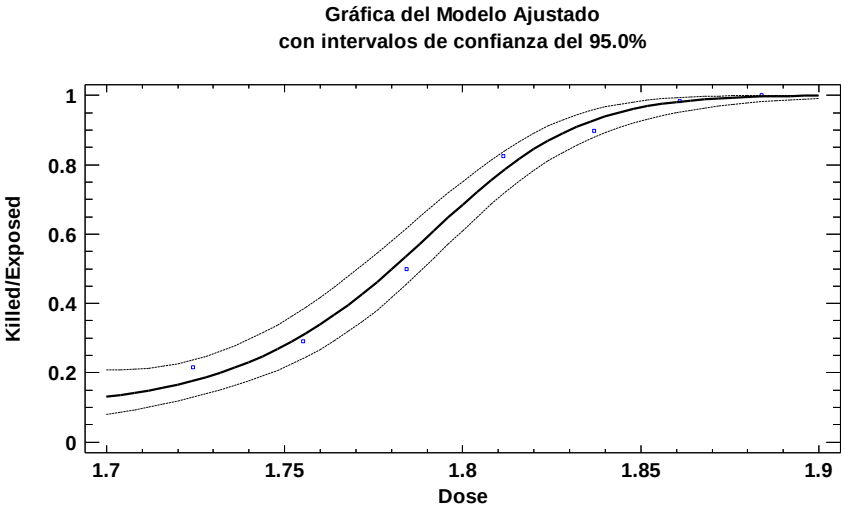
| *Fuente* | *Desviación* | *Gl* | *Valor-P* |
|---|---|---|---|
| Modelo | 281.008 | 2 | 0.0000 |
| Residuo | 3.19491 | 5 | 0.6700 |
| Total (corr.) | 284.202 | 7 | |

Porcentaje de desviación explicado por el modelo = 98.8758
Porcentaje ajustado = 96.7647

**Pruebas de Razón de Verosimilitud**

| *Factor* | *Chi-Cuadrada* | *Gl* | *Valor-P* |
|---|---|---|---|
| Dose | 7.04208 | 1 | 0.0080 |
| Dose^2 | 8.03733 | 1 | 0.0046 |

Ejercicio: Comprobar mediante el test de Wald y mediante el TRVG que el término cuadrático es significativo estadísticamente.




La pauta sistemática de variación de los residuos ha desaparecido

**Tabla de Predicciones Inversas para Dose**

| | | *LC Inferior 95.0%* | *LC Superior 95.0%* |
|---|---|---|---|
| *Porcentaje* | *Dose* | *Límite Conf.* | *Límite Conf.* |
| 40.0 | 1.76866 | 1.75774 | 1.77839 |
| 45.0 | 1.77475 | 1.76449 | 1.78389 |
| 50.0 | 1.78041 | 1.77075 | 1.78907 |
| 55.0 | 1.78581 | 1.77669 | 1.79408 |
| 60.0 | 1.79108 | 1.78247 | 1.79908 |

La $LD_{50}$ prevista es 1.78 con un intervalo de confianza (95%) [1.77 1.79]

## 11. PREDICCIÓN DE LA AFECTACIÓN DEL GANGLIO CENTINELA EN CÁNCER DE PRÓSTATA

En pacientes con cáncer de próstata es importante saber si el ganglio "centinela" está o no afectado. Con el fin de estudiar la posibilidad de evitar una laparotomía para averiguarlo, se estudió (Brown, 1980) si ciertas variables que podían ser conocidas sin cirugía (como el nivel de fosfatasa ácida en suero sanguíneo) permitían predecir la afectación del ganglio.

Las variables estudiadas fueron:

- Edad y nivel de fosfatasa ácida (variables cuantitativas continuas)
- "x-ray Reading" y "pathology reading" (grado) de una biopsia del tumor obtenido por punción antes de la cirugía (variables binarias con 2 alternativas 0 – 1)
- Medida aproximada del "estadío" ( 0 – 1) obtenida a partir de una palpación rectal

Se recogen a continuación los valores previos obtenidos en 53 pacientes, junto con el resultado ("Nodes"): ganglio afectado (1) o ganglio no afectado (0) encontrado en la posterior laparotomía. (Valores guardados en el archivo *prostata*)

| X-ray | Grade | Stage | Age | Acid | Nodes | X-ray | Grade | Stage | Age | Acid | Nodes |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 64 | 40 | 0 | 0 | 0 | 0 | 60 | 78 | 0 |
| 0 | 0 | 1 | 63 | 40 | 0 | 0 | 0 | 0 | 52 | 83 | 0 |
| 1 | 0 | 0 | 65 | 46 | 0 | 0 | 0 | 1 | 67 | 95 | 0 |
| 0 | 1 | 0 | 67 | 47 | 0 | 0 | 0 | 0 | 56 | 98 | 0 |
| 0 | 0 | 0 | 66 | 48 | 0 | 0 | 0 | 1 | 61 | 102 | 0 |
| 0 | 1 | 1 | 65 | 48 | 0 | 0 | 0 | 0 | 64 | 187 | 0 |
| 0 | 0 | 0 | 60 | 49 | 0 | 1 | 0 | 1 | 58 | 48 | 1 |
| 0 | 0 | 0 | 51 | 49 | 0 | 0 | 0 | 1 | 65 | 49 | 1 |
| 0 | 0 | 0 | 66 | 50 | 0 | 1 | 1 | 1 | 57 | 51 | 1 |
| 0 | 0 | 0 | 58 | 50 | 0 | 0 | 1 | 0 | 50 | 56 | 1 |
| 0 | 1 | 0 | 56 | 50 | 0 | 1 | 1 | 0 | 67 | 67 | 1 |
| 0 | 0 | 1 | 61 | 50 | 0 | 0 | 0 | 1 | 67 | 67 | 1 |
| 0 | 1 | 1 | 64 | 50 | 0 | 0 | 1 | 1 | 57 | 67 | 1 |
| 0 | 0 | 0 | 56 | 52 | 0 | 0 | 1 | 1 | 45 | 70 | 1 |
| 0 | 0 | 0 | 67 | 52 | 0 | 0 | 0 | 1 | 46 | 70 | 1 |
| 1 | 0 | 0 | 49 | 55 | 0 | 1 | 0 | 1 | 51 | 72 | 1 |
| 0 | 1 | 1 | 52 | 55 | 0 | 1 | 1 | 1 | 60 | 76 | 1 |
| 0 | 0 | 0 | 68 | 56 | 0 | 1 | 1 | 1 | 56 | 78 | 1 |
| 0 | 1 | 1 | 66 | 59 | 0 | 1 | 1 | 1 | 50 | 81 | 1 |
| 1 | 0 | 0 | 60 | 62 | 0 | 0 | 0 | 0 | 56 | 82 | 1 |
| 0 | 0 | 0 | 61 | 62 | 0 | 0 | 0 | 1 | 63 | 82 | 1 |
| 1 | 1 | 1 | 59 | 63 | 0 | 1 | 1 | 1 | 65 | 84 | 1 |
| 0 | 0 | 0 | 51 | 65 | 0 | 1 | 0 | 1 | 64 | 89 | 1 |
| 0 | 1 | 1 | 53 | 66 | 0 | 0 | 1 | 0 | 59 | 99 | 1 |
| 0 | 0 | 0 | 58 | 71 | 0 | 1 | 1 | 1 | 68 | 126 | 1 |
| 0 | 0 | 0 | 63 | 75 | 0 | 1 | 0 | 0 | 61 | 136 | 1 |
| 0 | 0 | 1 | 53 | 76 | 0 | | | | | | |

## Predicción de la afectación del ganglio centinela (continuación)

Con el fin de cuantificar el valor predictivo de las 5 variables explicativas consideradas se ajustó un MRL mediante la técnica “stepwise forward”.

Como valor límite del p-value para incluir variables en el modelo se adoptó 0.15, teniendo en cuenta el escaso número de observaciones disponible (sólo 20 casos positivos en la muestra).

Resultados obtenidos mediante Statgraphics:

Paso 0:
0 factores en el modelo. 52 g.l. para el error.
Porcentaje de desviación explicada = 0.00% Porcentaje ajustado = 0.00%

Paso 1:
Agregando factor X_ray con P-para-introducir = 0.000795212
1 factores en el modelo. 51 g.l. para el error.
Porcentaje de desviación explicada = 16.02% Porcentaje ajustado = 10.32%

Paso 2:
Agregando factor Estadio con P-para-introducir = 0.0174789
2 factores en el modelo. 50 g.l. para el error.
Porcentaje de desviación explicada = 24.05% Porcentaje ajustado = 15.51%

Paso 3:
Agregando factor fosf_acid con P-para-introducir = 0.100737
3 factores en el modelo. 49 g.l. para el error.
Porcentaje de desviación explicada = 27.89% Porcentaje ajustado = 16.50%

**Modelo Estimado de Regresión (Máxima Verosimilitud)**

| | | *Error* | *Razón de Momios* |
|---|---|---|---|
| *Parámetro* | *Estimado* | *Estándar* | *Estimada* |
| CONSTANTE | -3.57565 | 1.18097 | |
| Estadio | 1.75556 | 0.738918 | 5.7867 |
| fosf_acid | 0.0206294 | 0.0126482 | 1.02084 |
| X_ray | 2.06179 | 0.77758 | 7.86 |

**Análisis de Desviación**

| *Fuente* | *Desviación* | *Gl* | *Valor-P* |
|---|---|---|---|
| Modelo | 19.5925 | 3 | 0.0002 |
| Residuo | 50.6596 | 49 | 0.4079 |
| Total (corr.) | 70.2522 | 52 | |

Porcentaje de desviación explicado por el modelo = 27.8889
Porcentaje ajustado = 16.5013

Se constata que el modelo es globalmente muy significativo (p-value = 0.0002 correspondiente al valor 19.59 para una Gi-2 con 3 g.l.)

Por otra parte la deviance residual del modelo no es significativa (p-value = 0.4079 correspondiente al valor 50.66 para una Gi-2 con 49 g.l.)

Valor predictivo de las diferentes variables

La significación de los efectos de las 3 variables retenidas en el modelo, es decir la admisibilidad de la hipótesis $\beta_i = 0$ para cada variable, puede investigarse de dos formas alternativas:

- Test de Wald: dividiendo cada $b_i$ por su desviación típica y viendo el percentil correspondiente en una normal tipificada:

| | | |
|---|---|---|
| estadio: | 1.7556/0.7389 = 2.376 | p-value = 0.017 |
| fosf_acid: | 0.0206/0.0126 = 1.635 | p-value = 0.102 |
| X-ray: | 2.0618/0.776 = 2.651 | p-value = 0.008 |

-TRVG: estudiando la significación del incremento de la deviance residual cuando se elimina del modelo la variable en cuestión (es el procedimiento que sigue Stgraphics)

**Pruebas de Razón de Verosimilitud**

| *Factor* | *Chi-Cuadrada* | *Gl* | *Valor-P* |
|---|---|---|---|
| Estadio | 6.39907 | 1 | 0.0114 |
| fosf_acid | 2.69378 | 1 | 0.1007 |
| X_ray | 7.95386 | 1 | 0.0048 |

Por ambos caminos se obtiene que los efectos de Estadio y X-ray son claramente significativos.

La significación del efecto de la fosfatasa ácida es, sin embargo, dudoso (quizás debido a los pocos datos disponibles), puesto que el p-value correspondiente es 0.10

Estudio de la posible no linealidad del efecto de fosf_acid

(Nota: este estudio no aparece en el trabajo de Browm (1980) y lo hemos desarrollado como ejemplo para este curso)

Para contemplar la posibilidad de que el efecto de fosf_acid sobre el log-od ln(p/(1-p)) sea no lineal, se ha introducido en el modelo como variable explicativa adicional el cuadrado de fosf_acid. El ajuste resultante ha sido:

**Modelo Estimado de Regresión (Máxima Verosimilitud)**

| | | *Error* | *Razón de Momios* |
|---|---|---|---|
| *Parámetro* | *Estimado* | *Estándar* | *Estimada* |
| CONSTANTE | -8.63166 | 3.53479 | |
| Estadio | 1.58686 | 0.756508 | 4.88839 |
| X_ray | 2.07179 | 0.827636 | 7.93901 |
| fosf_acid | 0.144668 | 0.0845788 | 1.15566 |
| fosf_acid^2 | -0.000643497 | 0.000491403 | 0.999357 |

**Pruebas de Razón de Verosimilitud**

| *Factor* | *Chi-Cuadrada* | *Gl* | *Valor-P* |
|---|---|---|---|
| Estadio | 4.82007 | 1 | 0.0281 |
| X_ray | 7.10332 | 1 | 0.0077 |
| fosf_acid | 5.47225 | 1 | 0.0193 |
| fosf_acid^2 | 4.04521 | 1 | 0.0443 |

Se constata que en este nuevo modelo tanto fosf_acid (que mide el efecto sobre el log_odd cuando fosf_acid pasa de 0 a 1, como fosf_acid$^2$, que mide la curvatura de la relación entre el log-odd y la variable, son significativos.

Los signos obtenidos para ambos parámetros indican que el efecto sobre el log-odd del aumento de fosf_acid es más marcado para valores bajos de esta variable, siendo menos marcado a medida que ésta aumenta.

Este modelo presenta, sin embargo, el inconveniente de que el carácter parabólico de la relación propuesta entre el log-odd y la fosf_acid implica que a partir de un determinado valor de esta variable (concretamente 112), el log-odd disminuye al aumentar ésta, lo que no resulta lógico

Estudio de la posible no linealidad del efecto de fosf_acid (continuación)

Una posible alternativa para obviar este problema es incluir como variable explicativa en el MRL el log(fosf_acid) en vez de fosf_acid. En efecto, una relación lineal positiva entre Y y log(X), implica una relación no lineal entre Y y X con pendiente progresivamente menor pero que nunca llega a ser negativa

El ajuste de este nuevo modelo da los siguientes resultados, que ponen de manifiesto un efecto estadísticamente significativo (p-value=0.0366) de ln(fosf_acid) sobre el log-odd

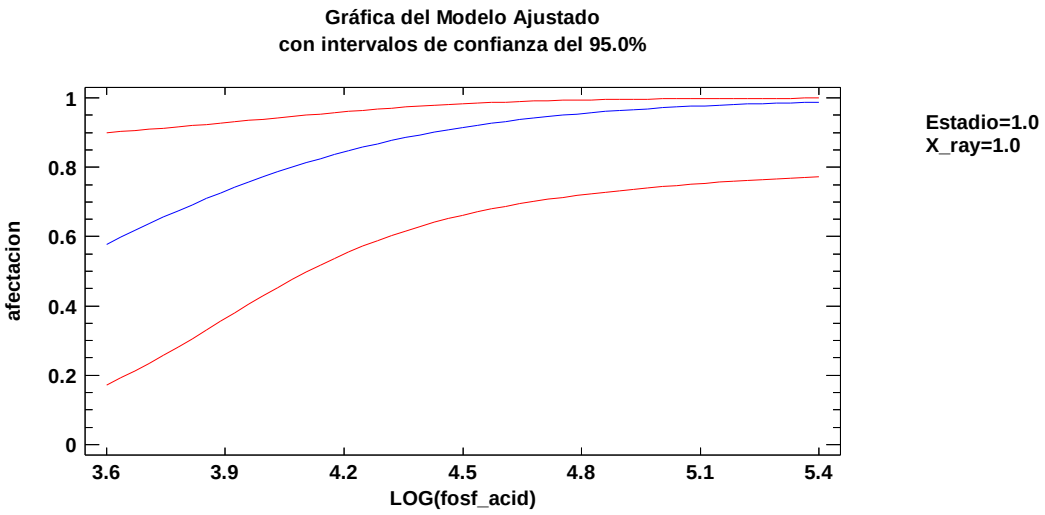
**Modelo Estimado de Regresión (Máxima Verosimilitud)**

| | | *Error* | *Razón de Momios* |
|---|---|---|---|
| *Parámetro* | *Estimado* | *Estándar* | *Estimada* |
| CONSTANTE | -11.7555 | 4.94966 | |
| Estadio | 1.76376 | 0.748257 | 5.83433 |
| X_ray | 2.05502 | 0.797625 | 7.80697 |
| LOG(fosf_acid) | 2.29224 | 1.13874 | 9.8971 |

**Pruebas de Razón de Verosimilitud**

| *Factor* | *Chi-Cuadrada* | *Gl* | *Valor-P* |
|---|---|---|---|
| Estadio | 6.28558 | 1 | 0.0122 |
| X_ray | 7.49832 | 1 | 0.0062 |
| LOG(fosf_acid) | 4.36741 | 1 | 0.0366 |

Para cuantificar el efecto de fosf_acid sobre la probabilidad de que el ganglio esté afectado, en el siguiente gráfica se ve la evolución de dicha probabilidad en función del valor de ln(fosf_acid), asumiendo que X-ray = 1 y Estadio =1




Se constata, por ejemplo, que en estas condiciones cuando fosf_acid pasa de 50 (log50=3.912) a 100 (log100=4.605), la probabilidad de que el ganglio esté afectado sube del 74% al 93% .

## 12. FACTORES PRONÓSTICO EN LA ENFERMEDAD HEMOLÍTICA NEONATAL

En un estudio sobre esta enfermedad se recogieron datos sobre 81 recién nacidos que la padecían, de los que 65 sobrevivieron. Las variables explicativas estudiadas fueron:

hemoglobina (**H**): concentración en el cordón umbilical (gr/100ml)

bilirrubina (**B**): concentración en sanfre (en mg/100ml)

Los datos, que se recogen en el archivo *enf_hemol_neonatal*, se ajustaron a un MRL obteniéndose los siguientes resultados

Variable dependiente: supervivencia (1 -SI 0 - NO)
Factores:
hemoglobina (gr/100ml en cordon umbilical)
bilirrubina (mg/100ml)

**Modelo Estimado de Regresión (Máxima Verosimilitud)**

| | | *Error* | *Razón de Momios* |
|---|---|---|---|
| *Parámetro* | *Estimado* | *Estándar* | *Estimada* |
| CONSTANTE | -2.53666 | 2.48701 | |
| hemoglobina | 0.547237 | 0.15747 | 1.72847 |
| bilirrubina | -0.482759 | 0.356676 | 0.617079 |

**Análisis de Desviación**

| *Fuente* | *Desviación* | *Gl* | *Valor-P* |
|---|---|---|---|
| Modelo | 40.1793 | 2 | 0.0000 |
| Residuo | 40.3283 | 78 | 0.9999 |
| Total (corr.) | 80.5076 | 80 | |

**Pruebas de Razón de Verosimilitud**

| *Factor* | *Chi-Cuadrada* | *Gl* | *Valor-P* |
|---|---|---|---|
| hemoglobina | 19.0786 | 1 | 0.0000 |
| bilirrubina | 1.88887 | 1 | 0.1693 |

Se constata que el efecto de **H** es muy significativo.

El efecto de **B** no llega a ser significativo, por lo que los autores decidieron no incluir esta variable en el modelo definitivo.

**Nota**: en nuestra opinión esta decisión es discutible, especialmente si el signo negativo obtenido para dicho efecto coincide con el que cabría esperar desde un punto de vista médico. Un p-value de 0.169 puede deberse a los pocos casos disponibles (sólo 16 fallecimientos) Además, en una situación como ésta, cometer un error de 1ª especie (rechazar la $H_0$ de que la bilirrubina no influye, cuando realmente es cierta), no tiene unas consecuencias tan graves como en otras situaciones, pues sólo introduciría un poco de “ruido” aleatorio en la ecuación de predicción.

## Predicción de la probabilidad de supervivencia

El modelo estimado, tras eliminar del mismo la variable **B** es

**Modelo Estimado de Regresión (Máxima Verosimilitud)**

| | | *Error* | *Razón de Momios* |
|---|---|---|---|
| *Parámetro* | *Estimado* | *Estándar* | *Estimada* |
| CONSTANTE | -5.49049 | 1.53297 | |
| hemoglobina | 0.636875 | 0.151581 | 1.89056 |

**Análisis de Desviación**

| *Fuente* | *Desviación* | *Gl* | *Valor-P* |
|---|---|---|---|
| Modelo | 38.2904 | 1 | 0.0000 |
| Residuo | 42.2172 | 79 | 0.9998 |
| Total (corr.) | 80.5076 | 80 | |

Se constata que la "deviance" residual no explicada por el modelo no es significativa

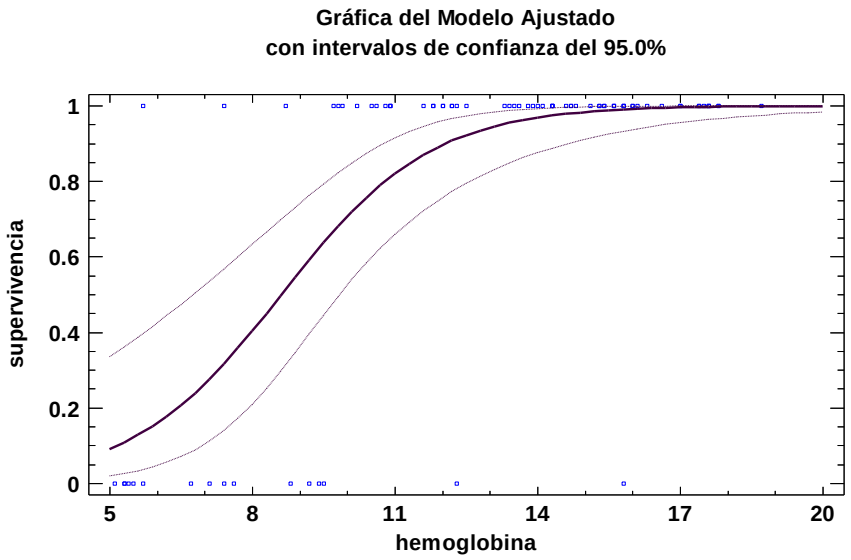
**Pruebas de Razón de Verosimilitud**

| *Factor* | *Chi-Cuadrada* | *Gl* | *Valor-P* |
|---|---|---|---|
| hemoglobina | 38.2904 | 1 | 0.0000 |

La ecuación final resultante para la probabilidad de supervivencia es

$$\text{Prob(Supervivencia)} = \frac{e^{-5.49+0.6369H}}{1+e^{-5.49+0.6369H}}$$

que se grafica a continuación, junto con el correspondiente intervalo de confianza para la predicción



# 13. MODELO DE REGRESIÓN MULTINOMIAL

## 13.1 Introducción

Una generalización del Modelo de Regresión Logística es el Modelo de Regresión Multinomial, que permite tratar situaciones en las que la variable respuesta tienen un número K, mayor que 2, de alternativas posibles.

Una posible forma de abordar este tipo de problemas consistiría en elaborar K modelos de regresión logística, uno para cada una de dichas alternativas

$$\mathrm{Prob}\left(Y = k / \vec{x}_j\right) = \frac{e^{\beta_{k0}+\beta_{k1}x_{1j}+\ldots+\beta_{kl}x_{lj}}}{1+e^{\beta_{k0}+\beta_{k1}x_{1j}+\ldots+\beta_{kl}x_{lj}}} \qquad \text{para } k = 1 \ldots K$$

Si embargo el ajuste independiente de estos K modelos no garantiza que para cualquier conjunto de valores $\vec{x}_j$ de las variables explicativas se verifique, tal como debería cumplirse, que $\sum_{k=1}^{K}\mathrm{Prob}\left(Y = k / \vec{x}_j\right) = 1$

## 13.2 Modelo de Regresión Logística Multinomial

Vamos a ver la forma de operar suponiendo que hay 3 alternativas (1, 2 y 3) para la variable respuesta Y, y un única variable explicativa x.

Sean $p_1 = P(Y=1)$ $p_2 = P(Y=2)$ $p_3 = P(Y=3)$

Se elige arbitrariamente una de las 3 alternativas como “base” (por ejemplo Y=3) y se definen dos modelos logísticos

$$\log\left(\frac{p_1}{p_3}\right) = \beta_{10} + \beta_{11}x \quad \rightarrow \quad \frac{p_1}{p_3} = e^{\beta_{10}+\beta_{11}x} \quad \rightarrow \quad p_1 = \left(e^{\beta_{10}+\beta_{11}x}\right)p_3$$

$$\log\left(\frac{p_2}{p_3}\right) = \beta_{20} + \beta_{21}x \quad \rightarrow \quad \frac{p_2}{p_3} = e^{\beta_{20}+\beta_{21}x} \quad \rightarrow \quad p_2 = \left(e^{\beta_{20}+\beta_{21}x}\right)p_3$$

Como $p_1 + p_2 + p_3 = 1$ se deduce inmediatamente las siguientes expresiones para las $p_i$ en función de los valores de los parámetros $\beta_{ij}$

$$p_1 = \frac{e^{\beta_{10}+\beta_{11}x}}{1+e^{\beta_{10}+\beta_{11}x}+e^{\beta_{20}+\beta_{21}x}}$$

$$p_2 = \frac{e^{\beta_{20}+\beta_{21}x}}{1+e^{\beta_{10}+\beta_{11}x}+e^{\beta_{20}+\beta_{21}x}}$$

$$p_3 = \frac{1}{1+e^{\beta_{10}+\beta_{11}x}+e^{\beta_{20}+\beta_{21}x}}$$

El ajuste simultáneo de las ecuaciones anteriores puede llevarse a cabo mediante diferentes software, incluyendo versiones recientes de SPSS.

## 13.3 Interpretación de los parámetros en el Modelo de Regresión Logística Multinomial

(Ejemplo ficticio)

Supongamos que se desea estudiar la influencia del sexo del niño sobre la aparición de ciertos problemas en el parto.

Se consideran sólo tres alternativas posibles para los partos:

Y = 1 Parto con el problema A

Y = 2 Parto con el problema B

Y = 3 Parto sin problemas

Se define para el sexo del nasciturus una variable binaria x, tal que x=0 si es niño y x=1 si es niña.

Se toma la tercera alternativa (parto sin problemas) como la de referencia

Operando con la primera de las ecuaciones se obtiene

$$\frac{p_1}{p_3} = \left(e^{\beta_{10}+\beta_{11}x}\right) = e^{\beta_{10}}e^{\beta_{11}x} \qquad \left(\frac{p_1}{p_3}\Big/\text{niño}\right) = e^{\beta_{10}} \qquad \left(\frac{p_1}{p_3}\Big/\text{niña}\right) = e^{\beta_{10}}e^{\beta_{11}}$$

Por lo tanto, en relación a la probabilidad de un parto normal, la probabilidad de la presencia del problema A se multiplica por $e^{\beta_{10}}$ en el caso de niños y por $e^{\beta_{10}}e^{\beta_{11}}$ en el caso de niñas.

Si, por ejemplo, $\beta_{11}$ resultara = 0, como $e^0=1$, el sexo no influiría sobre la probabilidad de que aparezca el problema A.

Razonando de forma análoga, si resultara $\beta_{21}>0$ se concluiría que la aparición del problema B es más probable en niñas que en niños.

## 13.4 Ejemplo: Efectividad de la estreptomicina en el tratamiento de la tuberculosis

En un estudio sobre el tema[6], 55 pacientes tuberculosos fueron tratados con estreptomicina y reposo mientras que otros 52 pacientes fueron tratados sólo con reposo.

Al cabo de 6 meses de tratamiento los pacientes fueron clasificados en 3 categorías[7]: Mejora, Igual, Empeora.

Los resultados obtenidos, junto con el sexo y la edad de los enfermos[8] se recogen en el archivo *tuberc.xls*

Los datos se ajustaron a un modelo de regresión logística multinomial, tomando la categoría "igual" como referencia

<u>Resultados del ajuste (utilizando SPSS)</u>

**Resumen del procesamiento de los casos**

| | | N | Porcentaje marginal |
|---|---|---|---|
| estado | igual | 22 | 20,6% |
| | mejor | 55 | 51,4% |
| | peor | 30 | 28,0% |
| Válidos | | 107 | 100,0% |
| Perdidos | | 0 | |
| Total | | 107 | |

[6] Medicla Research Council. *Streptomycin treatment of pulmonary tuberculosis.* Br. Med. J. 2:769 1948
[7] En el estudio se definieron 6 categorías, que aquí hemos agrupado para simplificar el ejemplo..
[8] Los datos de sexo y edad son ficticios, habiéndose generado al azar para jemplificar la posibilidad de varias variables explicativas

**Información del ajuste del modelo**

| Modelo | Criterio de ajuste del modelo | Contrastes de la razón de verosimilitud | | |
|---|---|---|---|---|
| | -2 log verosimilitud | Chi-cuadrado | gl | Sig. |
| Sólo la intersección | 188,942 | | | |
| Final | 172,169 | 16,773 | 6 | ,010 |

**Contrastes de la razón de verosimilitud**

| Efecto | Criterio de ajuste del modelo | Contrastes de la razón de verosimilitud | | |
|---|---|---|---|---|
| | -2 log verosimilitud del modelo reducido | Chi-cuadrado | gl | Sig. |
| Intersección | 173,821 | 1,652 | 2 | ,438 |
| estrept | 186,851 | 14,682 | 2 | ,001 |
| varon | 172,849 | ,680 | 2 | ,712 |
| edad | 173,756 | 1,587 | 2 | ,452 |

El estadístico de chi-cuadrado es la diferencia en las -2 log verosimilitudes entre el modelo final y el modelo reducido. El modelo reducido se forma omitiendo un efecto del modelo final. La hipótesis nula es que todos los parámetros de ese efecto son 0.

Los efectos de sexo y edad no resultan significativos, por lo que pueden eliminarse del modelo.

## Resultados incluyendo sólo la estreptomicina en el modelo

**Contrastes de la razón de verosimilitud**

| Efecto | Criterio de ajuste del modelo | Contrastes de la razón de verosimilitud | | |
|---|---|---|---|---|
| | -2 log verosimilitud del modelo reducido | Chi-cuadrado | gl | Sig. |
| Intersección | 16,610 | ,726 | 2 | ,695 |
| estrept | 30,400 | 14,516 | 2 | ,001 |

**Estimaciones de los parámetros**

| estado(a) | | B | Error típ. | Wald | gl | Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| | | Límite inferior | Límite superior | Límite inferior | Límite superior | Límite inferior | Límite superior |
| mejor | Intersección | ,125 | ,354 | ,125 | 1 | ,724 | |
| | estrept | 1,567 | ,543 | 8,328 | 1 | ,004 | 4,790 |
| peor | Intersección | ,288 | ,342 | ,709 | 1 | ,400 | |
| | estrept | ,069 | ,600 | ,013 | 1 | ,908 | 1,071 |

a La categoría de referencia es: igual.

Los dos modelos ajustados son:

$$\log\left(\frac{p_{mejora}}{p_{igual}}\right) = 0.125 + 1.567X_{estrept} \quad \rightarrow \quad p_{mejora} = p_{igual}e^{0.125+1.567X_{estrept}}$$

$$\log\left(\frac{p_{empeora}}{p_{igual}}\right) = 0.288 + 0.069X_{estrept} \quad \rightarrow \quad p_{empeora} = p_{igual}e^{0.288+0.069X_{estrept}}$$

Efecto de la estreptomicina sobre $p_{mejora}$

Del primer modelo se deduce:

- no utilizando estreptomicina: $p_{mejora} = p_{igual}e^{0.125} = p_{igual}x1.13$

- utilizando estreptomicina: $p_{mejora} = p_{igual}e^{0.125+1.567} = p_{igual}x5.44$

Sin utilizar estreptomicina $p_{mejora}$ es un 13% mayor que $p_{igual}$, mientras que utilizando estreptomicina $p_{mejora}$ es un 544% mayor que $p_{igual}$

El efecto de la estreptomicina sobre el logit de “mejora” frente a “igual” es muy significativo estadísticamente (valor_p = 0.004)

Sin embargo, el efecto de la estreptomicina sobre el logit de “empeora” frente a “igual” no es significativo estadísticamente (valor_p = 0.908)

Obtención de las probabilidades

Aplicando las expresiones

$$p_1 = \frac{e^{\beta_{10}+\beta_{11}x}}{1+e^{\beta_{10}+\beta_{11}x}+e^{\beta_{20}+\beta_{21}x}}$$

$$p_2 = \frac{e^{\beta_{20}+\beta_{21}x}}{1+e^{\beta_{10}+\beta_{11}x}+e^{\beta_{20}+\beta_{21}x}}$$

$$p_3 = \frac{1}{1+e^{\beta_{10}+\beta_{11}x}+e^{\beta_{20}+\beta_{21}x}}$$

es posible obtener las probabilidades de las tres alternativas según se utilice o no la estreptomicina.

Probabilidades utilizando estreptomicina (x=1)

Probabilidad de “igual”

$$p_{igual} = \frac{1}{1+e^{\beta_{10}+\beta_{11}x1}+e^{\beta_{20}+\beta_{21}x1}} = \frac{1}{1+e^{0.125}e^{1.567}+e^{0.288}e^{0.069}} = \frac{1}{1+1.133x4.792+1.334x1.071} = 0.127$$

Probabilidad de “mejora”

$$p_{mejora} = p_{igual}e^{0.125+1.567} = p_{igual}x4.23 = 0.127x5.44 = 0.691$$

Probabilidad de “empeora”

$$p_{empeora} = p_{igual}e^{0.288+0.069} = p_{igual}x1.43 = 0.127x1.43 = 0.182$$

Como es lógico se cumple que $p_{igual} + p_{mejora} + p_{empeora} = 1$

El SPSS calcula las probabilidades previstas por el modelo y las compara con las frecuencias relativas realmente observadas en los datos.

**Frecuencias observadas y pronosticadas**

| | | Frecuencia | | | Porcentaje | |
|---|---|---|---|---|---|---|
| estrept | estado | Observada | Pronosticada | Residuo de Pearson | Observada | Pronosticada |
| 0 | igual | 15 | 15,000 | ,000 | 28,8% | 28,8% |
| | mejor | 17 | 17,000 | ,000 | 32,7% | 32,7% |
| | peor | 20 | 20,000 | ,000 | 38,5% | 38,5% |
| 1 | igual | 7 | 7,000 | ,000 | 12,7% | 12,7% |
| | mejor | 38 | 38,000 | ,000 | 69,1% | 69,1% |
| | peor | 10 | 10,000 | ,000 | 18,2% | 18,2% |

Los porcentajes se basan en las frecuencias observadas totales de cada subpoblación.

En este caso las probabilidades previstas por el modelo coinciden exactamente con las frecuencias relativas observadas en los datos.

**Nota:**

La coincidencia es lógica, puesto que sólo hay 4 probabilidades independientes a las que ajustar el modelo (las 3 probabilidades cuando x=0 y las 3 probabilidades cuando x=1 deben sumar 1, por lo que conocidas 2 de cada grupo las terceras se obtendrían por diferencia) y el modelo tiene exactamente 4 parámetros