Regresión Logística

Introducción

En este tema estudiaremos cómo construir y analizar un *modelo de regresión* que pretende representar la dependencia lineal de una **variable respuesta con dos categorías** (dicotómica) respecto a otras variables explicativas (categóricas o cuantitativas). Representaremos (sin pérdida de generalidad) las dos posibles respuestas de la variable respuesta como **0** y **1**. *Por ejemplo queremos explicar la probabilidad de que una persona enferme (1) o no (0) en función de su edad, sexo, hábitos relacionados con el alcohol,…*

Se dice que un proceso es **binomial** cuando sólo tiene dos posibles resultados: "**éxito**" (representado por el valor "**1**") y "**fracaso**" (representado por el valor "**0**"), siendo la probabilidad de cada uno de ellos constante en una serie de repeticiones.

Un proceso binomial está caracterizado por la probabilidad de éxito, representada por **p** (es el único parámetro de su función de probabilidad), la probabilidad de fracaso es por tanto **1-p** ya que, evidentemente, ambas probabilidades deben sumar 1. En ocasiones, se usa el cociente $\frac{p}{1-p}$, denominado ***"odds"***, y que indica cuánto más probable es el éxito que el fracaso.

En los modelos de Regresión Logística se **pretende estudiar si la probabilidad de éxito (p) de una variable de este tipo depende, o no, de otra u otras variables.**

Modelo de Regresión Logística Simple

Tal y como hemos comentado nuestro objetivo es estudiar la relación de la probabilidad de "éxito" en nuestro proceso con una serie de variables explicativas. Como primera modelización del problema con un modelo de regresión lineal nos podríamos plantear el modelo:

$$p = \alpha_0 + \alpha_1 X$$

Siendo **p** la probabilidad de éxito y **X** la variable explicativa (que supondremos cuantitativa).

El hecho de que el valor de **p** deba estar necesariamente entre 0 y 1 (puesto que es una probabilidad) nos supone el primer problema, ya que si la variable explicativa puede tomar valores en un rango amplio nos va a ser "muy difícil" determinar los coeficientes del modelo (α_0, α_1) de forma que **p** no salga del rango (0,1).

Por este motivo no se modeliza directamente "p", sino $\frac{p}{1-p}$ (odds), y como este cociente tiene el problema añadido de estar acotado teniendo que ser necesariamente mayor que 0, por comodidad y para trabajar con toda la recta real modelizaremos $\log(\frac{p}{1-p})$, que puede tomar cualquier valor de la recta real.

Así, el modelo de regresión sería de la forma:

$$\log(\frac{p}{1-p}) = \alpha_0 + \alpha_1 X$$

Que es equivalente a la expresión:

$$\Pr(\text{Re}\,spuesta) = p = \frac{\exp(\alpha_0 + \alpha_1 X)}{1 + \exp(\alpha_0 + \alpha_1 X)}$$ **(Función de distrib. Logística)**

Pero, con esta modelización **¿que significado tendrán los *coeficientes del modelo* (α_0, α_1)?**

$$\frac{p}{1-p} = \exp(\alpha_0 + \alpha_1 X) \quad \Rightarrow \quad \frac{p}{1-p} = \exp(\alpha_0) * \exp(\alpha_1 X)$$

Así, **exp(α_0)** representaría el valor del ODDS cuando la variable explicativa toma el valor 0, es decir, cuánto más probable es el éxito que el fracaso cuando la variable explicativa vale 0.

$$\frac{p}{1-p} = \exp(\alpha_0 + \alpha_1 X) \text{ y } \frac{p'}{1-p'} = \exp(\alpha_0 + \alpha_1 (X+1)) \Rightarrow \exp(\alpha_1) = \frac{\frac{p'}{1-p`}}{\frac{p}{1-p}} \text{ =OR}$$

Por lo que **exp(α_1)** representa el **OR** (**Odds Ratio**) por unidad de incremento de la variable explicativa X.

La **estimación de los coeficientes**, aunque existen otros métodos, se suele realizar por el más extendido que es el de **máxima verosimilitud**, que consiste en maximizar la función de verosimilitud de la muestra. Pero estos cálculos los realizan los paquetes de software estadísticos, por lo que **nosotros únicamente nos debemos centrar en la interpretación de los resultados**.

Ejemplo 1: Efecto dosis-respuesta de un tratamiento sobre la curación

Se desea estudiar el efecto **dosis-respuesta** para un tratamiento midiendo la respuesta "curación" o "no curación". El objetivo es medir cómo influyen las distintas dosis del tratamiento sobre la probabilidad de curación. A una muestra aleatoria de enfermos se la divide también aleatoriamente en 4 grupos en los que se administra el tratamiento según la siguiente tabla:

Grupo	Tratamiento
1	0 mg (No tratamiento)
2	50 mg
3	100 mg
4	150 mg

Se mide la respuesta como **curación** o **no curación**.

La variable explicativa *X* tiene 4 valores que pueden ser, bien los mg de cada dosis (0, 50, 100, 150) o bien un código arbitrario para cada dosis (p.e. 0, 1, 2, 3). Tanto una codificación como la otra para la variable explicativa puede considerarse en este caso cuantitativa: la primera porque directamente contiene los mg de cada tratamiento, y la segunda porque recoge de un valor al siguiente saltos equidistantes de 50 mg.

Se plantea un modelo logístico para modelizar (o explicar) la probabilidad de curación en función de la dosis administrada.

¿Cómo interpretaríamos los coeficientes?:

- α_0 será el logaritmo del ***odds*** de la dosis 0 (el logaritmo del *odds* de la curación para los enfermos no tratados). O lo que es lo mismo, **exp(α_0)** representa *cuánto más probable es la curación frente a la no curación en enfermos no tratados (dosis=0).*
- α_1 será el logaritmo del ***odds ratio*** por aumento de unidad de dosis. O lo que es lo mismo, **exp(α_1)** representa el Odds Ratio al incrementar una **unidad** la dosis, *incremento de la probabilidad de curación al incrementar una **unidad** la dosis*.

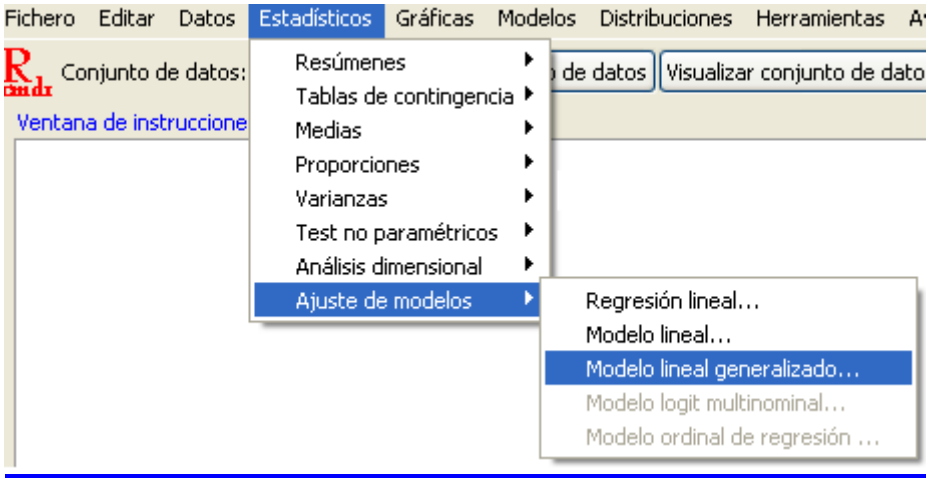
Nota.- Conviene tener presentes dos aspectos:

1.- La dependencia de α_1 de la codificación de la variable *X*: si se usan los mg la **unidad** es 1 mg y si se usan los códigos, la **unidad** es el cambio de dosis (de un grupo a otro la diferencia es 50 mg).

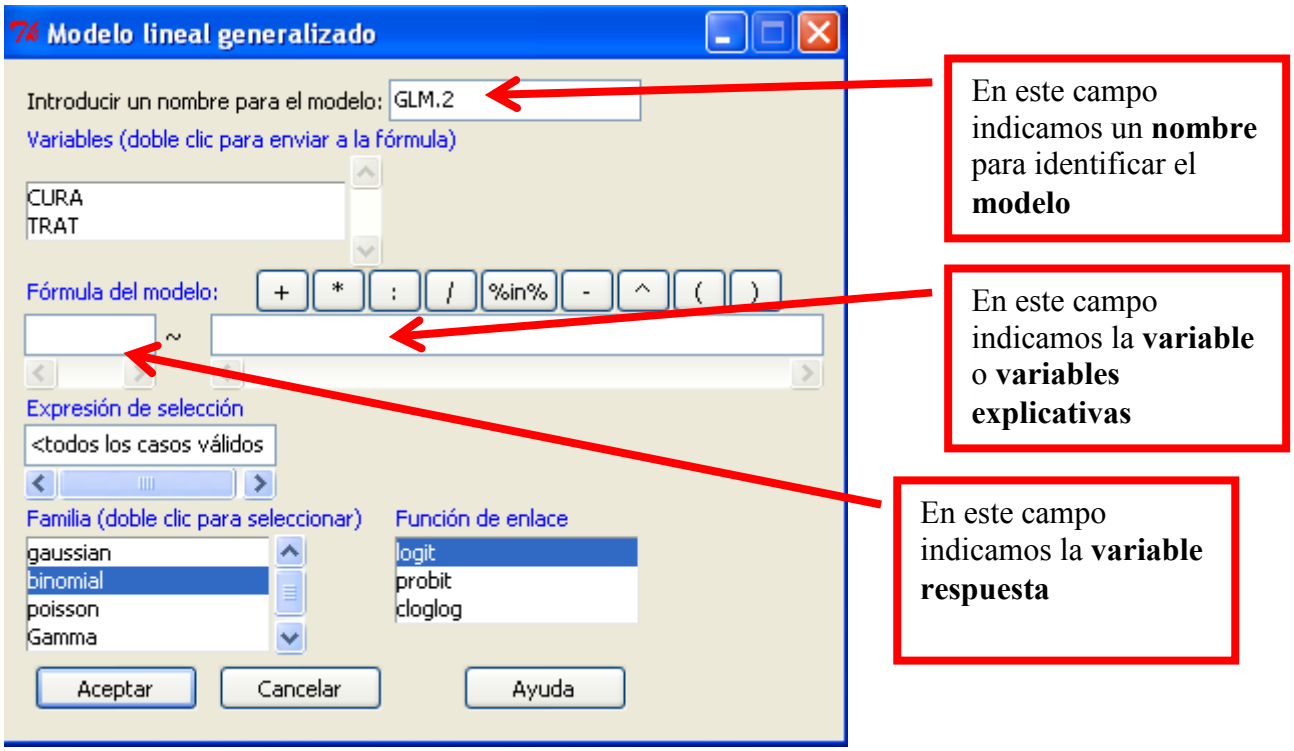
2.- El modelo asume que el cambio en el logaritmo del odds es constante, es decir el logaritmo del odds cambia α_1 por pasar de la dosis 0 a la 1, o por pasar de la dosis 2 a la 3.

Regresión logística en R-Commander.

Para plantear y resolver un modelo de regresión logística mediante el software R-Commander podemos acudir a la opción ***Estadísticos>Ajuste de modelos>Modelo lineal generalizado:***



Aparecerá la siguiente ventana:



Por último pulsamos **aceptar** (dejamos las opciones que nos da la aplicación por defecto, es decir, la familia **binomial** y la función **logit** son correctas)

Ejemplo 2: Eficacia de dos tratamientos sobre la curación

Se quiere comparar la eficacia de dos tratamientos alternativos para una misma enfermedad. Asumiendo que la respuesta que vamos a medir sólo tiene dos

posibles resultados: **curación** o **no curación** y que la probabilidad de curación es la misma para todos los enfermos, se trata de un proceso binomial.

Objetivo: Estudiar si este proceso curación/no curación está asociado, o no, con el tratamiento, es decir, si la probabilidad de curación dado el tratamiento *A* es igual, o distinta, a la probabilidad de curación dado el tratamiento *B*.

Supóngase que sobre una muestra aleatoria de 40 enfermos, dividida aleatoriamente en dos grupos de 20, a cada uno de los cuales se le suministra un tratamiento, se obtienen los siguientes resultados:

	Tratamiento A (X=1)	**Tratamiento B (X=0)**
Curación	18	13
no curación	2	7
Total	**20**	**20**

Si se define la variable tratamiento como X=1 para el tratamiento A y X=0 para el tratamiento B, a partir de la tabla podemos estimar la probabilidad de curación para el tratamiento B: $p|(X=0)=13/20$ y para el tratamiento A: $p|(X=1)=18/20$ Como ambas probabilidades son distintas, "parece" que la probabilidad de curación depende del tratamiento. Las preguntas son: ¿esta dependencia es generalizable ("estadísticamente significativa")? ¿cuánto depende ("clínicamente relevante")?
La primera pregunta la podemos resolver mediante la prueba χ^2, la segunda mediante las denominadas "medidas de asociación" como "*odds ratio*" (OR).

En el ejemplo: OR: ((18/20)/(2/20))/(13/20)/(7/20) =(18x7)/(13x2) = 4,85

En caso de no diferencia OR vale 1. Recordemos que el OR es la medida más extendida y que es conveniente que a estas estimaciones puntuales las

acompañemos de su intervalo de confianza que nos indica la precisión de la estimación.

Para realizarlo con un paquete estadístico hay que partir de un archivo en que los datos estén *individualizados*, es decir un archivo con 40 casos (los enfermos) con dos variables una para el tratamiento con los valores 0 y 1 y otra para el resultado, también con dos valores 0: no curación y 1: curación. Sería, por tanto:

Curación	Tratamiento	
1	0	
.	.	**13 casos**
1	0	
0	0	
.	.	**7 casos**
0	0	
1	1	
.	.	**18 casos**
1	1	
0	1	
0	1	**2 casos**

y parte del resultado del procesamiento mediante el paquete estadístico R-Commander (**Opción de menú: *Estadísticos>Ajuste de modelos>Modelo lineal generalizado***) se muestra a continuación:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.6190     0.4688   1.320    0.187
TRAT          1.5782     0.8805   1.792    0.073
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretación de los resultados "***Coefficients***":

- La columna **`Estimate`** es la estimación de los coeficientes (`TRAT` para la estimación del coeficiente de variable Tratamiento α_1 y `Intercept` para la estimación de α_0)
- La columna `Pr(>|z|)` es el p-valor de los contrastes correspondientes a la significatividad de los coeficientes:

Fila `TRAT`	Fila `Intercept`
H_0: $\alpha_0=0$ H_1: $\alpha_0\neq 0$	H_0: $\alpha_1=0$ H_1: $\alpha_1\neq 0$
p-valor=0.073	p-valor=0.187

Indican en el caso de la fila *Constante* si el valor de α_0 es significativamente diferente de 0 o no, o lo que es lo que sería lo mismo, si $\exp(\alpha_0)$ es significativamente diferente de 1 o no. El que $\exp(\alpha_0)$ sea igual a 1 representaría que la probabilidad de curación es igual a la de no curación para X=0 (es decir para tratamiento 2)

Y en la fila TRAT si el valor de α_1 es significativamente diferente de 0 o no, o lo que sería lo mismo, si $\exp(\alpha_1)$ es significativamente diferente de 1 o no. El que $\exp(\alpha_1)$ sea igual a 1 representaría que el Odds Ratio de pasar de tratamiento 2 a tratamiento 1 es 1, es decir que no hay diferencias en cuanto a probabilidad de curación entre los tratamientos 1 y 2.

- Obteniendo $\exp(\alpha_0)=\exp(0.6190)=1.8570$, podemos interpretar que en Tratamiento 2 la probabilidad de curación es 1.857 mayor que la probabilidad de no curación.
- Obteniendo $\exp(\alpha_1)=\exp(1.5782)=4.85$, podemos interpretar que la probabilidad de curación es aprox. 4.85 veces mayor en el grupo con Tratamiento 1 que en grupo con Tratamiento 2. (Nótese que la estimación del OR coincide con las obtenidas anteriormente)

Para obtener los intervalos de confianza de las estimaciones de los coeficientes de este modelo, una vez obtenidos los resultados del modelo seleccionamos la opción de menú: ***Modelos>Intervalos de confianza*** y obtendremos los resultados:

```
> Confint(GLM.2, level=.95, type="LR")
                 2.5 %   97.5 %
(Intercept) -0.27345789 1.597783
TRAT        -0.01772915 3.590490
```

Para los extremos de los intervalos de confianza también debemos aplicar la función exponencial:

(Intercept)

exp(-0.27345789)= 0.7607444

exp(1.597783)= 4.942064

TRAT

exp(-0.01772915) = 0.982427

exp(3.590490)= 36.25183

Así, podemos resumir en una tabla los resultados obtenidos mediante el R-Commander:

Resumen de los resultados obtenidos por este modelo:.

(Tal cual nos lo da R-Commander)

Parámetro	**Estimación**	**Int.Conf.95%**	**p-valor**
α_0	0.6190	(-0.2735, 1.5978)	0.187
α_1	1.5782	(-0.0177, 3.5905)	0.073

(Aplicando la función exponencial para obtener estimaciones directas del Odds y OR)

Parámetro	Estimación	Int.Conf.95%	p-valor
exp(α_0)=Odds de curación cuando X=0 (es decir, Tratamiento 2)	1.8570	(0.7607 , 4.9421)	0.187
exp(α_1)=OR de X=1 frente X=0 (es decir, Tratamiento 1 frente a Tratamiento 2)	4.85	(0.9824 , 36.2518)	0.073

Es equivalente interpretar el p-valor del contraste del OR anterior y observar el intervalo de confianza que se muestra. Si el intervalo de confianza al 95% contiene el valor 1, el contraste anterior no será rechazado con el nivel de significatividad igual a 0.05 (es decir, el p-valor será mayor que 0.05), y si por el contrario el intervalo de confianza al 95% no contiene el valor 1 el contraste anterior será rechazado con el nivel de significatividad 0.05 al obtener un p-valor menor que este valor.

Modelo de Regresión Logística Múltiple

Es una generalización del Modelo Simple considerando en esta ocasión como variables explicativas las variables $X_1, X_2, ..., X_k$:

$$\ln\left(\frac{p}{1-p}\right) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + ... + \alpha_k X_k$$

Y la interpretación de los coeficientes es también una generalización, es decir, e^{α_0} es el *odds* cuando todas las X_i=0 y e^{α_i} es el *odds ratio* por el aumento de una unidad en la variable X_i manteniendo constantes las otras (*controlando por ellas*). Cuando no tiene sentido físico X_i=0, e^{α_i} se interpreta como el *odds* basal, es decir, el *odds* que no depende de las variables independientes.

Los coeficientes se estiman y los contrastes de hipótesis se realizan del mismo modo que en el modelo simple, aunque con el modelo múltiple (igual que en

regresión lineal) se pueden hacer contrastes no sólo sobre cada coeficiente, sino también sobre el modelo completo o para comparar modelos

Variables Indicadoras ("dummy")

En los modelos de Regresión Logística, al igual en Regresión Lineal, en la modelización se asume que las variables son "**cuantitativas**". Si una variable categórica se introduce en el análisis con las categorías {0,1,2,3}, la interpretación del coeficiente que acompañará a la misma indicará que para la categoría "2" el efecto es el doble que para la categoría codificada como "1", cuando esto puede no tener ningún sentido (pensemos, por ejemplo, que "0" representa '0 gramos', "1" representa '100 gr', "2" representa '165 gr' y "3" representa '235 gr')

La solución es la misma que en regresión lineal; crear tantas variables como categorías menos 1 denominadas ***variables indicadoras*** o ***dummy*** con el siguiente esquema:

	X_1	X_2	X_3
Dosis 0	0	0	0
Dosis 1	1	0	0
Dosis 2	0	1	0
Dosis 3	0	0	1

El modelo quedaría $\ln(\frac{p}{1-p}) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \ldots$

¿Cómo interpretamos los coeficientes?

- Para la dosis 0, como las tres variables son 0:

$\ln(\frac{p}{1-p} \mid Dosis = 0) = \alpha_0$, es decir $\exp(\alpha_0)$ es el *odds* para la dosis 0

- Para la dosis 1 el modelo queda:

$\ln(\frac{p}{1-p} \mid Dosis = 1) = \alpha_0 + \alpha_1$, por lo tanto $\exp(\alpha_1)$ es el OR de la dosis 1 con respecto a la dosis 0, del mismo modo $\exp(\alpha_2)$ es el OR de la dosis 2 con respecto a la dosis 0, etc.

Conviene destacar que estas variables indicadoras no tienen ningún sentido por sí solas y por tanto deben figurar en los modelos y se debe contrastar su inclusión siempre en bloque.

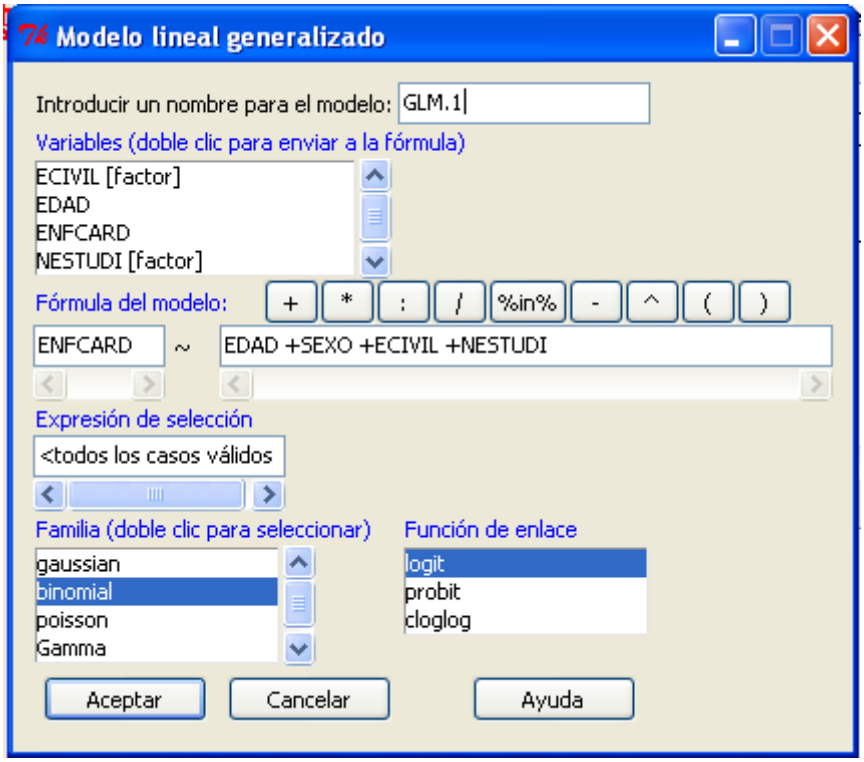
Ejemplo 3: Asociación entre diferentes variables explicativas y la Enfermedad Cardiovascular.

Se quiere relacionar la presencia o no de Enfermedad cardiovascular en personas que tienen una edad igual o superior a 30 años con una serie de variables que se detallan a continuación:

- ENFCARD: Enfermedad cardiovascular codificada como “Sí”, “No”
- SEXO: Codificada como “Mujer” , “Hombre”
- EDAD
- ECIVIL: Estado civil codificado como “0=Soltero, 1=Casado, 2=Viudo, 3=Separado/Divorciado ,4= No Sabe/No contesta”
- NESTUDI: Nivel de estudios codificado como “0=Sin estudios, 1=Analfabeto, 2=Estudios primarios, 3=Bachillerato, 4=Licenciatura, 5=No sabe/No contesta”

Para estudiar estas relaciones planteamos un modelo de Regresión Logística cuya variable respuesta será “ENFCARD” y el resto serán variables explicativas en el modelo.

Con este fin seleccionamos la opción de menú ***Estadísticos>Ajuste de modelos>Modelo lineal generalizado*** en el que indicamos:



Una vez seleccionamos la variable a explicar (ENFCARD) y las variables explicativas (EDAD, SEXO, ECIVIL, NESTUDI) pulsamos **Aceptar**.

Los resultados obtenidos son:

```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -10.12479    2.31809  -4.368 1.26e-05 ***
EDAD                            0.25380    0.06029   4.210 2.56e-05 ***
SEXO[T.Mujer]                  -0.02301    0.59564  -0.039    0.969
ECIVIL[T.Casado]               -0.81676    1.25142  -0.653    0.514
ECIVIL[T.Viudo]                 0.10938    1.23058   0.089    0.929
ECIVIL[T.Separado - Divorciado]  13.82553 2218.47725   0.006    0.995
ECIVIL[T.No Sabe - No Contesta]  -1.76026    1.90230  -0.925    0.355
NESTUDI[T.Analfabeto]           12.49753 2747.98114   0.005    0.996
NESTUDI[T.Estudios Primarios]    0.17422    1.98269   0.088    0.930
NESTUDI[T.Bachillerato]          0.87386    1.09746   0.796    0.426
NESTUDI[T.Diplomatura]           1.39097    1.38580   1.004    0.316
NESTUDI[T.Licenciatura]          0.50952    1.18319   0.431    0.667
NESTUDI[T.No sabe-No contesta]   1.22754    1.20951   1.015    0.310
```

Y a continuación podemos pedir también los intervalos de confianza para las estimaciones en ***Modelos>Intervalos de confianza:***

```
> Confint(GLM.14, level=.95, type="LR")
                                        2.5 %     97.5 %
(Intercept)                       -15.3864114 -6.2028941
EDAD                                0.1531320  0.3917764
SEXO[T.Mujer]                      -1.1866005  1.1769997
ECIVIL[T.Casado]                   -3.3498124  1.6500648
ECIVIL[T.Viudo]                    -2.3925994  2.5609728
ECIVIL[T.Separado - Divorciado] -108.7046194         NA
ECIVIL[T.No Sabe - No Contesta]    -5.7104756  2.2494562
NESTUDI[T.Analfabeto]            -147.8407474         NA
NESTUDI[T.Estudios Primarios]      -3.5328455  4.2789581
NESTUDI[T.Bachillerato]            -1.3102343  3.1080908
NESTUDI[T.Diplomatura]             -1.3860807  4.1713298
NESTUDI[T.Licenciatura]            -1.8437693  2.8896530
NESTUDI[T.No sabe-No contesta]     -1.1503557  3.6741292
```

Vamos a analizar los resultados obtenidos para cada una de las variables una a una:

- Interceptación del modelo α_0

```
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -10.12479    2.31809   -4.368 1.26e-05 ***
```

exp(α_0)=exp(-10.12479)= 0.00005 sería el Odds de enfermedad cuando todas las variables tuvieran el valor "0" (no tiene sentido su interpretación, puesto que la variable EDAD=0 no tiene sentido). No se interpreta pero es necesario incluirla en el modelo (***).

- Edad:

```
                              Estimate Std. Error z value Pr(>|z|)
EDAD                          0.25380    0.06029  -4.210 2.56e-05 ***
```

exp(α_i)=exp(0.2538)= 1.29 sería el OR por cada año que pasa. Es decir, una persona que tiene (X+1) años tiene 1.29 veces más probabilidad de desarrollar la enfermedad que una persona que tiene (X) años. Y esta relación es significativa, lo que indica el p-valor menor que 0.05 y la marca ***. Concretamente, el intervalo de confianza al 95% para este OR es: (exp(0.1531) , exp(0.3918))=(1.1654, 1.4796), es decir, una persona que tiene (X+1) años tiene entre 1.1654 y 1.4796 veces más probabilidad de desarrollar la enfermedad que una persona que tiene (X) años con un 95% de confianza.

- Sexo:

	Estimate	Std. Error	z value	Pr(>|z|)
SEXO[T.Mujer]	0.02301	0.59564	0.039	0.969

exp(α_i)=exp(0.02301)= 1.0233 sería el OR de la categoría "Mujer" frente a la otra, es decir, frente a Hombre. La estimación indica que las mujeres tiene 1.02 veces más probabilidad de tener una enfermedad cardiovascular que los hombres, pero esta asociación no es significativa, puesto que el p-valor (0.969) es mucho mayor que 0.05. Si calculáramos el intervalo de confianza para este parámetro obtendríamos un intervalo que contiene el valor 1 y que por tanto indica que la relación no es significativa: (exp(-1.1866) , exp(1.1769))=(0.3052, 3.2446).

- Estado civil:

	Estimate	Std. Error	z value	Pr(>|z|)
ECIVIL[T.Casado]	-0.81676	1.25142	-0.653	0.514
ECIVIL[T.Viudo]	0.10938	1.23058	0.089	0.929
ECIVIL[T.Separado - Divorciado]	13.82553	2218.47725	0.006	0.995
ECIVIL[T.No Sabe - No Contesta]	-1.76026	1.90230	-0.925	0.355

En este caso se muestran las estimaciones que nos permiten obtener los OR de cada categoría de Estado Civil frente a la primera categoría (que no aparece): "soltero". En cualquier caso estas asociaciones no son significativas puesto que los p-valores que se obtienen en todas ellas son muy superiores al nivel de significatividad estándar (0.05).

- Nivel de estudios:

Lo mismo ocurre con esta variable, los resultados muestran las estimaciones que permiten obtener los OR de cada categoría de esta variable frente a la primera categoría "Sin estudios". En este caso tampoco aparece ninguna relación significativa, puesto que todos los p-valores son inferiores a 0.05.

Interacción y confusión en la regresión logística

Cualquier modelo de regresión puede tener dos objetivos:

1) ***predictivo,*** en el que el interés del investigador es predecir lo mejor posible la variable dependiente, usando un conjunto de variables explicativas.

2) ***estimativo,*** en el que el interés se centra en estimar la relación de una o más variables independientes con la variable dependiente. El segundo objetivo es el más frecuente en estudios etiológicos en los que se trata de encontrar factores determinantes de una enfermedad o un proceso.

La ***interacción*** y la ***confusión*** son dos conceptos importantes cuando se usan los modelos de regresión con objetivo estimativo, que tienen que ver con la interferencia que una o varias variables pueden realizar en la asociación entre otras.

Confusión:

Existe **confusión** cuando la asociación entre dos variables difiere significativamente según se considere, o no, otra variable, a esta última variable se le denomina *variable de confusión* para la asociación.

El modelo más sencillo para estudiar la asociación entre una variable binomial y otra variable X_1 es "$\log(\frac{p}{1-p}) = \alpha_0 + \alpha_1 X_1$" donde α_1 cuantifica la asociación.

Se dice que X_2 es una variable de confusión para esta asociación, si el modelo "$\log(\frac{p}{1-p}) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2$" produce una estimación para α_1 diferente del modelo anterior. Evidentemente esta definición se puede ampliar a un conjunto de variables, se dice que las variables $X_2, ..., X_k$ son variables de confusión si la estimación de α_1 obtenida por el modelo "$\log(\frac{p}{1-p}) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + ... + \alpha_k X_k$" es diferente de la obtenida en el modelo simple. En ambos casos se dice que la estimación de α_1 obtenida en los modelos múltiples está ***controlada*** o ***ajustada*** por X_2 o por $X_2, ..., X_k$.

Interacción:

Existe **interacción** cuando la asociación entre dos variables varía según los diferentes niveles de otra u otras variables.

El modelo más sencillo que hace explícita la interacción entre dos variables X_1 y X_2 es "$\log(\frac{p}{1-p}) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_1 X_2$."

Si α_3 es distinto de 0, el *odds ratio* por una unidad de cambio en X_1 manteniendo fijo X_2 es diferente para cada valor x_2 de X_2. Del mismo modo, el *odds ratio* por una unidad de cambio en X_2 manteniendo fijo X_1 es diferente para cada valor x_1 de X_1.

Es obvio que primero debe contrastarse la interacción y después, en caso de que no exista, la confusión.

Estrategias de modelización

Debido a los dos objetivos distintos que un análisis de regresión puede tener es difícil establecer una estrategia general para encontrar el mejor modelo de regresión, es más, el mejor modelo significa cosas distintas con cada objetivo.

En un análisis predictivo el mejor modelo es el que produce predicciones más *fiables* para una nueva observación, mientras que en un análisis estimativo el mejor modelo es el que produce estimaciones más *precisas* para el coeficiente de la variable de interés.

- ***Especificación del modelo máximo***

Se trata de establecer todas las variables que van a ser consideradas. El modelo máximo deberá tener menos variables independientes que el número de datos menos uno (n-1) (un criterio habitual es incluir como máximo una variable cada 10 eventos).

El criterio para decidir qué variables forman el modelo máximo lo establece el investigador en función de sus objetivos y del conocimiento teórico que tenga

sobre el problema, evidentemente cuanto menor sea el conocimiento previo mayor tenderá a ser el modelo máximo.

En un modelo máximo grande aumenta la probabilidad de ***problemas de colinealidad***.

En el modelo máximo deben considerarse también los términos de interacción que se van a introducir (en un modelo estimativo sólo interesan interacciones entre la variable de interés y las otras)

- ***Comparación de modelos***

Debe establecerse cómo y con qué se comparan los modelos. Si bien hay varios estadísticos sugeridos para comparar modelos, el más frecuentemente usado es el ***logaritmo del cociente de verosimilitudes***, recordando que cuando los dos modelos sólo difieren en una variable, el contraste con el logaritmo del cociente de verosimilitudes es equivalente al contraste de Wald, pero a veces interesa contrastar varias variables conjuntamente mejor que una a una (por ejemplo todos los términos no lineales) o, incluso, es necesario hacerlo (por ejemplo para variables indicadoras).

Los distintos modelos a comparar se pueden construir de dos formas: por eliminación o hacia atrás ("***backward***") y por inclusión o hacia adelante ("***forward***").

Con la primera estrategia, se ajusta el modelo máximo (con todas las variables) y se calcula el logaritmo del cociente de verosimilitudes para cada variable como si fuera la última introducida, se elige el menor de ellos (el que indica menor relación con la vaiable) y se contrasta con el nivel de significación elegido. Si es mayor o igual que el valor crítico se adopta este modelo como resultado del análisis y si es menor se elimina esa variable y se vuelve a repetir todo el proceso hasta que no se pueda eliminar ninguna variable.

Con la estrategia hacia adelante, se empieza con un modelo de una variable, aquella que presente el mejor logaritmo del cociente de verosimilitudes (la que presente indicios de mayor relación). Se calcula el logaritmo del cociente de verosimilitudes para la inclusión de todas las demás, se elige el menor de ellos y se contrasta con el nivel de significación elegido. Si es menor que el valor crítico, se para el proceso y se elige el modelo simple como mejor modelo, y si

es mayor o igual que dicho valor crítico, esa variable se incluye en el modelo y se vuelve a calcular el logaritmo del cociente de verosimilitudes para la inclusión de cada una de todas las restantes, y así sucesivamente hasta que no se pueda incluir ninguna más.

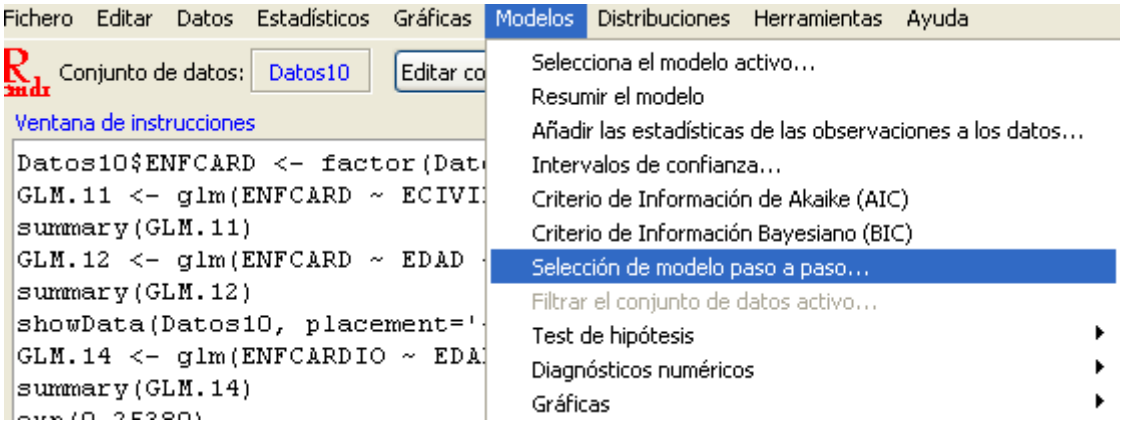
Una modificación de esta última estrategia es la denominada "***stepwise***" que consiste en que, cada vez que con el criterio anterior se incluye una variable, se calculan los logaritmos del cociente de verosimilitudes de todas las incluidas hasta ese momento como si fueran las últimas y la variable con menor logaritmo del cociente de verosimilitudes no significativo, si la hubiera, se elimina. Se vuelven a calcular los logaritmos del cociente de verosimilitudes y se continua añadiendo y eliminando variables hasta que el modelo sea estable.

Cuando se contrasta un término de interacción, el modelo debe incluir todos los términos de orden inferior y, si como resultado del contraste, dicho término permanece en el modelo, también ellos deben permanecer en el mismo, aunque no se pueda rechazar que los coeficientes correspondientes no son distintos de cero.

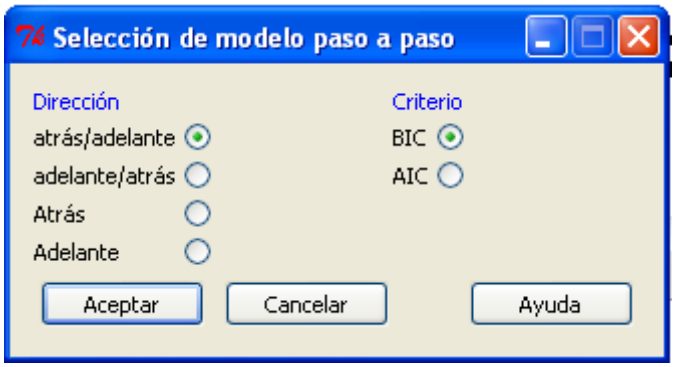
Una esopción que puede ayudar a buscar el mejor modelo: Seleccionar el modelo por más de uno de los procesos comentados, y si el modelo final coincide en todos ellos es muy probable que sea el mejor seleccionado.

Continuación del Ejemplo 3: Asociación entre diferentes variables explicativas y la Enfermedad Cardiovascular.

En este ejemplo disponíamos de 4 variables para explicar la presencia de enfermedad cardiovascular. Para seleccionar el mejor modelo (si es el que contiene solo la edad, o la edad y el sexo, o cualquier otra combinación de variables) podemos acudir en el menú (y tras haber pedido el modelo completo tal y como se ha indicado en la introducción del ejemplo 3) a: ***Modelos>Selección de modelo paso a paso***



Aparecerá la ventana:



En la que podemos seleccionar cuatro formas de seleccionar el mejor modelo (Dirección) y dos criterios diferentes que miden la fuerza de cada relación (Criterio).

Por todos los métodos y criterios se selecciona para este caso el modelo que únicamente contiene la EDAD para explicar la presencia de ENFERMEDAD CARDIOVASCULAR, indicando que el resto de variables no tiene una relación significativa con esta enfermedad.

La colinealidad en regresión logística

Al igual que en Regresión Lineal Múltiple la existencia de colinealidad entre las variables explicativas puede producir complicaciones en la estimación de los efectos de las variables influyentes. Es aconsejable estudiar los posibles problemas de colinealidad entre las variables explicativas, aunque durante la estimación del modelo se pueden detectar problemas de este tipo cuando, por

ejemplo, una variable con efecto significativo en el modelo deja de serlo cuando se introduce otra variable que tampoco resulta significativa. Si ambas tienen problemas de colinealidad, el modelo puede no distinguir a cuál de las dos se debe el efecto apareciendo por tanto ambas como no significativas. En este caso, la eliminación de una de ellas podría producir un empeoramiento en la bondad de ajuste del modelo.

En cualquier caso, los problemas de colinealidad y su control y estudio son exactamente los mismos que en Regresión Lineal Múltiple, por lo que se pueden aplicar los mecanismos ya vistos en esa sección.

Evaluación de los modelos de regresión logística

Una vez encontrado el mejor modelo, hay que validarlo, es decir ver si "trabaja" igual con otros individuos distintos de aquellos con los que se ha generado.

En un *modelo estimativo* se trata de ver si se obtiene el mismo *odds ratio* para la variable de interés. Aquí nos vamos a enfocar en los *modelos predictivos* en los que validar significa ver si el modelo predice bien la variable dependiente en un nuevo individuo. Ello implica dos conceptos relacionados validez (*"accuracy"*) y generalizabilidad (*"generalizability"*).

La **validez** es el grado en que las predicciones coinciden con las observaciones y tiene dos componentes: *calibración* y *discriminación*.

-La ***calibración*** compara el número predicho de eventos con el número observado en grupos de individuos,

- La ***discriminación*** evalúa el grado en que el modelo distingue entre individuos en los que ocurre el evento y los que no.

Por ejemplo, en un modelo logístico para predecir muerte en la UCI, si la mortalidad observada en la muestra es 27%, el modelo estará perfectamente calibrado si predice una mortalidad de 27%, sin embargo podría no distinguir entre los pacientes que mueren y los que sobreviven. A la inversa, si el modelo asignara una probabilidad de muerte de 2% a todos los pacientes que sobreviven y una probabilidad de 4% a todos los que mueren, el modelo tendría una perfecta discriminación, pero estaría pobremente calibrado.

La **generalizabilidad** es la capacidad del modelo de realizar predicciones válidas en individuos diferentes de aquellos en los que se ha generado y tiene también dos componentes: *reproducibilidad y transportabilidad*

- La ***reproducibilidad*** es la capacidad del modelo de realizar predicciones válidas en individuos no incluidos en la muestra con la que se ha generado, pero procedentes de la misma población)
- La ***transportabilidad*** es la capacidad de realizar predicciones válidas en pacientes procedentes de una población distinta pero relacionada.

La prueba estadística que evalúa la calibración es la de **Hosmer-Lemeshow**, aplicada sobre la misma muestra de trabajo (validez interna) o sobre la muestra, o el grupo, de validación (generalizabilidad).

Como medida de discriminación se utiliza el **área bajo la curva ROC** que representa para todos los pares posibles de individuos formados por un individuo en el que ocurrió el evento y otro en el que no, la proporción de los que el modelo predice una mayor probabilidad para el que tuvo el evento. A partir de un área de 0,7 la discriminación del modelo se considera aceptable.

Referencias específicas:

http://www.hrc.es/bioest/M_docente.html

V. Abraira, A.Pérez de Vargas- Métodos Multivariantes en Bioestadística.
Ed. Centro de Estudios Ramón Areces. 1996.

L.C. Silva Ayçaguer - Excursión a la regresión logística en Ciencias de la Salud Díaz de Santos. 1995

D.W. Hosmer, S. Lemeshow - Applied Logistic Regression. John Wiley & Sons. 1989.

Assessing the generalizability of prognostic information. Justice AC. et al. *Ann Intern Med*. 130: 515-524. 1999.

RESOLUCIÓN DE CASOS PRÁCTICOS:

1. EJERCICIO PRÁCTICO 1: Regresión Logística Simple

En un estudio para ver la dependencia de la dosis en el efecto de un veneno, se seleccionan aleatoriamente 4 grupos de 4 animales (12 animales en total) cada uno a los que se suministran distintas dosis (0, 1, 2, 3) de veneno y se observan las muertes provocadas. Las dosis equivalen a:

Dosis	**Mg**
0	0 mg
1	0.2 mg
2	0.4 mg
3	0.6 mg

Los resultados resumidos se muestran en la siguiente tabla:

Dosis	**0**	**1**	**2**	**3**
Animales	*4*	*4*	*4*	*4*
Muertes	**0**	**1**	**3**	**3**

ACTIVIDADES:

a) Introduce los datos en el R-Commander (Nota- Tendrías que tener dos columnas, “Dosis (0,1,2 o 3)” y “Muerte(0=”no”,1=”sí”) y 12 filas, una por cada animal)

b) Realiza una Regresión Logística para estudiar el efecto de la dosis sobre la probabilidad de muerte.

c) Analiza los resultados que proporciona este paquete estadístico y da una explicación sobre la interpretación de los coeficientes del modelo ajustado.

2. EJERCICIO PRÁCTICO 2: Regresión Logística Múltiple.

Se desea estudiar, mediante un modelo de regresión logística, la posible asociación entre el cáncer de vejiga y las siguientes variables explicativas:

- Consumo habitual de café (1=Sí/0=No)
- Ambiente de residencia (1=Urbana/0=Rural)
- Edad
- Consumo habitual de tabaco (1=Sí/0=No)
- Sexo (1=Mujer/0=Hombre)
- Consumo habitual de alcohol (1=Sí/0=No)

Se dispone de datos de 100 personas, 50 pacientes con cáncer y 50 individuos sin la enfermedad cuyos datos se guardan en la variable CÁNCER con los valores 0 (no cáncer) y 1 (cáncer).

ACTIVIDADES:

a) Los datos en formato SPSS se encuentran en el fichero “cancer.sav”. En primer lugar debes importarlos al R-Commander para comenzar con el análisis.

b) Realiza el análisis mediante el ajuste de un modelo de Regresión Logística, estudiando los efectos simples y la posible existencia de interacciones entre las variables.

c) Selecciona el modelo que mejor ajusta los datos disponibles y da una interpretación de los coeficientes del mismo.