Departamento de Estadística
Universidad Carlos III de Madrid

BIOESTADISTICA (55 - 10536)

Introducción a la regresión logística

1. INTRODUCCIÓN

La regresión logística es un procedimiento cuantitativo de gran utilidad para problemas donde la variable dependiente toma valores en un conjunto finito. Su uso se impone de manera creciente desde la década de los 80 debido a las facilidades computacionales con que se cuenta desde entonces. A continuación, desarrollaremos el caso especial en que la variable dependiente o respuesta es dicotómica.

Supongamos que la variable dependiente Y representa la ocurrencia o no de un suceso, por ejemplo:

- un paciente muere o no antes del alta.
- una persona deja o no de fumar después de un tratamiento.
- en un estudio retrospectivo un individuo es caso o control.
- un paciente positivo al VIH está o no en el estado IV.

Podemos decir que la variable dependiente Y toma valor 1 si ocurre el suceso, y valor 0 si no ocurre el suceso.

Por otra parte nos interesa estudiar la relación entre una o más variables independientes o explicativas: X_1, X_2, ..., X_p y la variable Y. El modelo logístico establece la siguiente relación entre la probabilidad de que ocurra el suceso, dado que el individuo presenta los valores $X_1=x_1$, $X_2=x_2$,...,$X_p=x_p$:

$$\Pr(Y=1\,|\,x_1,x_2,\cdots,x_p)=\frac{1}{1+\exp(-\alpha-\beta_1x_1-\beta_2x_2-\cdots-\beta_px_p)}.$$

Otra forma de presentar esta relación es:

$$\operatorname{logit}(\Pr(Y=1\,|\,x))=\log\left(\frac{\Pr(Y=1\,|\,x)}{1-\Pr(Y=1\,|\,x)}\right)=\alpha+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p,$$

donde denotamos con $\Pr(Y=1\,|\,x)$ la probabilidad condicional $\Pr(Y=1\,|\,x_1,x_2,\cdots,x_p)$.

Un problema importante es estimar los parámetros α, β_i's, a partir de un conjunto de observaciones. El procedimiento de estimación de estos parámetros se basa en el método de máxima verosimilitud. Existen varios programas que realizan estas estimaciones, por ejemplo: LOGIT, RELODI (que utilizaremos en nuestra exposición), MULTLR, EPISTAT, BMDP, SAS, etc., mediante la obtención del máximo del logaritmo de la función de verosimilitud:

$$L(y,\beta)=\sum_{i=1}^{n}y_i\ln(p_i)+(1-y_i)\ln(1-p_i),$$

donde n es el número de observaciones y $p_i=\Pr(Y=y_i\,|\,x_i)$.

Una vez que hayamos calculado los estimadores máximo-verosímiles (MV) de β_i's, puede interesarnos el cálculo de intervalos de confianza de estos parámetros, para ello podemos utilizar la estimación de la matriz de covarianza de los estimadores MV de los β_i. El intervalo de confianza del 100*(1-α)% puede calcularse por:

$$\hat{\beta}_i \pm z_{1-\alpha/2}\sqrt{\hat{\mathrm{Var}}(\hat{\beta}_i)}..$$

Podemos también contrastar la hipótesis nula H_0:β_i=0 mediante el siguiente estadístico: $Z = \dfrac{\hat{\beta}_i}{\sqrt{\hat{\mathrm{Var}}(\hat{\beta}_i)}}$.

Otra vía para probar la hipótesis anterior, cuando se consideran varias variables, es utilizando el máximo de la función de verosimilitud. Ejemplificaremos el procedimiento para el caso de dos variables X_1 y X_2. Se consideran los siguientes modelos:

$$\text{Modelo 1: } \mathrm{logit}(\Pr(Y=1|X_1)) = \alpha + \beta_1 X_1.$$
$$\text{Modelo 2: } \mathrm{logit}(\Pr(Y=1|X_2)) = \alpha + \beta_2 X_2.$$
$$\text{Modelo 3: } \mathrm{logit}(\Pr(Y=1|X_1,X_2)) = \alpha + \beta_1 X_1 + \beta_2 X_2.$$

Nos interesa en el modelo 3 probar las hipótesis H_0:β_1=0 y H_0:β_2=0. Sean $\hat{L}_1$, $\hat{L}_2$ y $\hat{L}_3$ los máximos de la función de verosimilitud para los modelos 1, 2 y 3, respectivamente. Se cumple que: $-2\ln(\hat{L}_2) - 2\ln(\hat{L}_3) \approx Z^2$ donde $Z = \dfrac{\hat{\beta}_1}{\sqrt{\hat{\mathrm{Var}}(\hat{\beta}_1)}}$, o sea, el estadístico para la primera de las hipótesis.

De manera análoga se tiene: $-2\ln(\hat{L}_1) - 2\ln(\hat{L}_3) \approx Z^2$ con $Z = \dfrac{\hat{\beta}_2}{\sqrt{\hat{\mathrm{Var}}(\hat{\beta}_2)}}$.

EJEMPLO: En una sala de terapia se desea estudiar la relación entre la sobrevivencia y las variables edad e infarto agudo del miocardio. A continuación mostramos los resultados del programa RELODI para datos de 200 pacientes tratados en esa sala.

Salida abreviada de RELODI (Modelo 1)

```
Número de casos para los cuales FALLECIDO es igual a 1: 76
Tamaño total de la muestra 200

-2 ln verosimilitud final: 245.91

Coeficiente   S.E.         z-score
-2.1920
 0.0373       0.0094       3.8009 EDAD
```

Salida abreviada de RELODI (Modelo 2)

```
Número de casos para los cuales FALLECIDO es igual a 1: 76
Tamaño total de la muestra 200

-2 ln verosimilitud final: 260.64

Coeficiente   S.E.         z-score
-0.6931
 0.2531       0.2954        0.8509 INFARTO
```

Salida abreviada de RELODI (Modelo 3)

```
Número de casos para los cuales FALLECIDO es igual a 1: 76
Tamaño total de la muestra 200

-2 ln verosimilitud final: 244.29

Coeficiente   S.E.         z-score
-2.4340
 0.0370       0.0094        3.8973 EDAD
 0.3935       0.3112       1.2645 INFARTO
```

Comprobemos las fórmulas aproximadas:

$$-2\ln(\hat{L}_2) - 2\ln(\hat{L}_3) = 260.64 - 244.29 = 16.35 \approx Z^2 = 3.8973^2 \approx 15.19.$$

$$-2\ln(\hat{L}_1) - 2\ln(\hat{L}_3) = 245.91 - 244.29 = 1.62 \approx Z^2 = 1.2645^2 \approx 1.60.$$

Notemos que este enfoque nos permite probar hipótesis del tipo: H_0:ß$_{p+1}$=0, ß$_{p+2}$=0, ..., ß$_{p+q}$=0 en el modelo: $\text{logit}(\Pr(Y=1|x)) = \alpha + \sum_{i=1}^{p+q} \beta_i x_i$ bastará calcular el máximo de la función de verosimilitud para este modelo y para el modelo siguiente: $\text{logit}(\Pr(Y=1|x)) = \alpha + \sum_{i=1}^{p} \beta_i x_i$. Se utiliza el siguiente estadístico:

$$\chi^2 = -2\ln\left(\frac{\hat{L}_p}{\hat{L}_{p+q}}\right),$$

donde $\hat{L}_{p+q}$ es el máximo de la función de verosimilitud para el primer modelo y $\hat{L}_p$ es el máximo para la función de verosimilitud del segundo modelo. El estadístico, bajo la hipótesis nula, se distribuye como una χ^2_q.

Para evaluar el grado de concordancia entre los valores observados de *Y*, y los valores estimados de $\hat{p}$ se puede utilizar el siguiente estadístico de bondad de ajuste: $\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{p}_i)^2}{\hat{p}_i(1-\hat{p}_i)}$. Esta medida es inestable para valores de $\hat{p}$ cercanos a 0 ó a 1.

2. EJEMPLOS DE USOS DE LA REGRESIÓN LOGÍSTICA

Estudios Descriptivos:

La regresión logística puede utilizarse como método descriptivo cuando se desea estudiar desde una perspectiva epidemiológica la aparición de un determinado evento en un grupo de individuos, por ejemplo:

- los pacientes de una determinada enfermedad desarrollan un cierto signo propio de ésta.
- los niños dejan la lactancia materna exclusiva.
- el fallecimiento de individuos de una cohorte.

EJEMPLO: Se seleccionan al azar *n* (300) historias clínicas de enfermos de la patología en estudio, se determina la fecha de detección de la enfermedad t_d, si el paciente tiene el signo de interés se toma la fecha en que apareció t_s, si el paciente no tiene el signo se toma la fecha de la última consulta t_e. Con estos datos definimos la variable dependiente *Y* como 1 si el paciente no tiene el signo, y como 0 si lo tiene, y la variable independiente *t* como la diferencia en días de la fecha de aparición y la detección en

caso de que Y=0 o la diferencia de la fecha de la última anotación en la historia clínica y la fecha de detección si Y=1, o sea: $t = \begin{cases} t_s - t_d & si\ Y = 0 \\ t_e - t_d & si\ Y = 1 \end{cases}$. Se ajusta el siguiente modelo:

$$\Pr(Y = 1 \mid t) = \frac{1}{1 + \exp(-\alpha - \beta t)}.$$

Salida Abreviada de RELODI

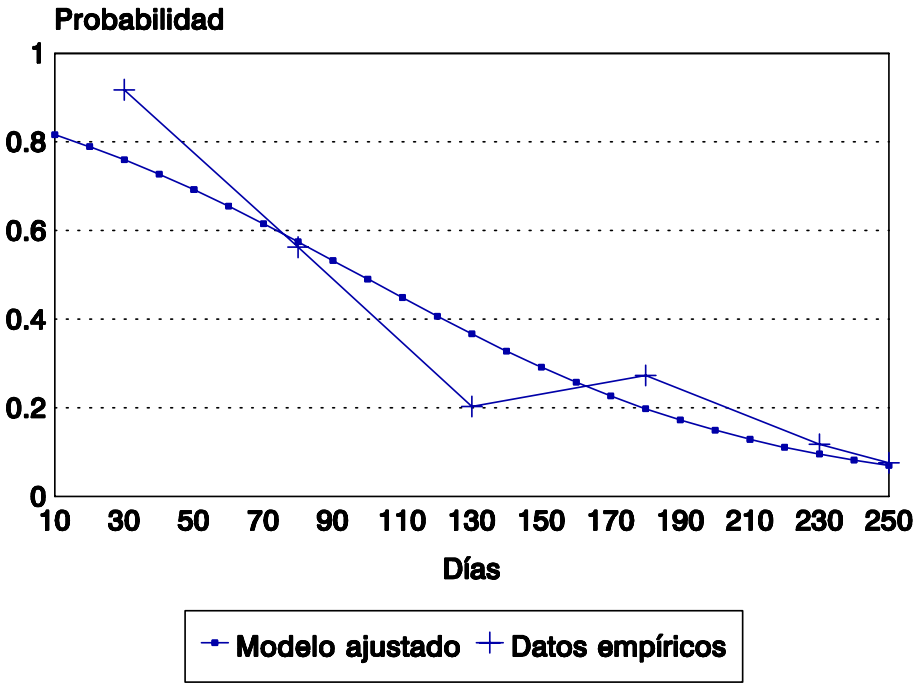
```
Número de casos para los cuales SIGNO es igual a 1: 110
Tamaño total de la muestra 300

Coeficiente   S.E.          z-score
 1.6642
-0.0168       0.0021        -7.7585 TIEMPO
```

Se tiene entonces que $\alpha \approx 1.664$ y $\beta \approx -0.017$. Por tanto, la probabilidad de que un paciente no tenga el signo a t días de la detección de la enfermedad se estima por:

$$\Pr(Y = 1 \mid t) = \frac{1}{1 + \exp(-1.664 + 0.017t)}.$$

De esta manera podemos calcular $\Pr(Y = 1 \mid t)$ para distintos valores de la variable t = 20, 40, 60,..., esta probabilidad no es más que la prevalencia de pacientes que a t días no tienen el signo en estudio. En el figura 1 se presenta la curva de prevalencia estimada por el modelo. Si el ajuste de la curva es adecuado los datos empíricos (proporción de individuos sin el signo en un grupo de estudio cuya variable t esté en un rango predefinido), serán cercanos a la curva teórica.

Figura 1. Prevalencia de No signo después de la detección.



Modelo Estadístico de Pronóstico:

Si se desea estimar la probabilidad de la ocurrencia de un suceso en función de un grupo de variables explicativas (predictoras) conocidas: X_1, X_2, ..., X_p, puede fijarse un modelo logístico, una vez que se hayan estimado los parámetros α y ß$_i$'s, puede calcularse la probabilidad $\Pr(Y=1|x_1,x_2,\cdots,x_p)$ para cualquier individuo cuyos variables independientes toman valores: x_1, x_2, ..., x_p, respectivamente.

Ejemplos de este tipo de estudios se presentan en la siguiente tabla:

Suceso a predecir	Variables predictoras
El tiempo de duración de la estancia de una hospitalización es superior a 7 días	Edad, sexo, diagnóstico principal, procedimiento quirúrgico principal, hospitalizaciones anteriores
Sobrevivencia de un paciente que ingresa a un servicio de quemados	Edad, porcentaje de quemaduras de primer y segundo grado, es o no diabético
Un niño padece de parasitismo intestinal	Edad, lugar de residencia, estatura, peso, resultados académicos
El tiempo de sobrevivencia de una paciente que ha sido operada de cáncer de mama es superior a 5 años	Edad de la paciente al momento de la operación, año calendario de la operación, número de nódulos positivos detectados

EJEMPLO: Se desea conocer la probabilidad de que un paciente que se ingresa en una sala de terapia intensiva sobreviva. Para este tipo de estudios es recomendable la definición de grupos diagnósticos (conjunto de entidades o enfermedades que tienen en común afectar a un mismo sistema del organismo), por tanto en nuestro ejemplo nos limitaremos a algunas de las variables que puedan influir el pronóstico de la evolución de pacientes con Enfermedades del Sistema Cardiocirculatorio (ESCC):

Edad (años) X_1
Enfermedad Hipertensiva (S/N) X_2
Insuficiencia Cardiaca (S/N) X_3
Disrritmia (S/N) X_4
Infarto Agudo del Miocardio (S/N) X_5
Enfermedad Pulmonar Obstructiva Crónica y afecciones afines (S/N) X_6
Ingresos anteriores por estas causas (#) X_7

Se estudiarán entonces un grupo de pacientes que ingresen a la sala de terapia intensiva con diagnóstico de ESCC, se les medirán las variables anteriores, que definiremos como 1 si hay presencia del problema y como 0 si no. Se espera entonces al egreso de cada paciente, si egresa vivo la variable Y toma valor 1, en caso contrario toma valor 0. La matriz de los datos de este estudio puede ser, por ejemplo:

Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	51	0	0	0	0	0	0
1	54	0	0	0	1	0	1
:	:	:	:	:	:	:	:
0	46	0	1	1	0	0	0

Se ajusta el siguiente modelo: $\Pr(Y=1|x_1,x_2,...,x_7)=\dfrac{1}{1+\exp(-\alpha-\beta_1x_2-\beta_2x_2-...-\beta_7x_7)}$.

Salida Abreviada de RELODI

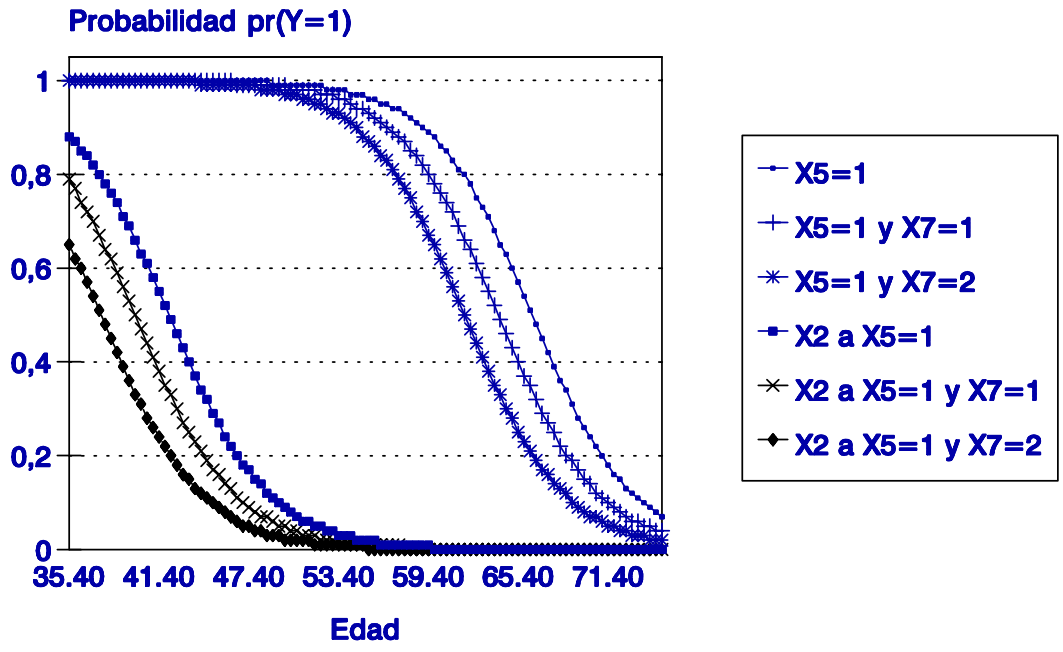
```
Número de casos para los cuales VIVO es igual a 1: 100
Tamaño total de la muestra 200

Coeficiente   S.E.          z-score
22.2266
-0.3115       0.0782        -3.9837 EDAD
-1.3663       0.6175        -2.2126 HIPERTENSION
-3.3569       0.7488        -4.4826 INSUFICIENCIA
-2.5825       0.6801        -3.7970 DISRRITMIA
-2.2972       0.6823        -3.3668 INFARTO
 0,3243       0.6401         0.5066 EPOC
-0.6813       0.2066        -3.2968 INGRESOS
```

Dados los parámetros estimados la probabilidad de sobrevivencia $\Pr(Y=1|x_1, x_2, ..., x_7)$ está dada por:

$$\Pr(Y=1|x_1, x_2, ..., x_7) = \frac{1}{1+\exp(-22.2+0.3x_1+1.4x_2+3.4x_3+2.6x_4+2.3x_5-0.3x_6+0.7x_7)}.$$

En la figura 2 se presentan distintas curvas de sobrevivencia utilizando el modelo logístico anterior.

Figura 2. Curvas de Sobrevivencia (Modelo Logístico)



Notemos como disminuye la probabilidad de sobrevivencia con la edad, con la cantidad de ingresos previos y la conjunción de varias patologías.

De esta misma manera podemos contemplar variables referentes a procedimientos terapéuticos, determinándose cuales son mejores (ofrezcan una mayor probabilidad de sobrevivencia) según las condiciones del paciente.

Análisis de Factores de Riesgo:

La regresión logística puede utilizarse como método para la estimación de la razón de disparidad (odds ratio *OR*). Veamos como obtenemos el *OR* en el caso de una variable independiente *X*, tenemos que: $OR = \frac{\Pr(Y=1|X=1)\Pr(Y=0|X=0)}{\Pr(Y=0|X=1)\Pr(Y=1|X=0)}$, y si asumimos el siguiente modelo: $\text{logit}(\Pr(Y=1|X)) = \ln\left(\frac{\Pr(Y=1|X)}{\Pr(Y=0|X)}\right) = \alpha + \beta X$ que para X=1 y X=0 toma las siguientes expresiones: $\ln\left(\frac{\Pr(Y=1|X=1)}{\Pr(Y=0|X=1)}\right) = \alpha + \beta$ y $\ln\left(\frac{\Pr(Y=1|X=0)}{\Pr(Y=0|X=0)}\right) = \alpha$, de donde obtenemos, $\ln(OR) = \ln\left(\frac{\Pr(Y=1|X=1)}{\Pr(Y=0|X=1)}\frac{\Pr(Y=0|X=0)}{\Pr(Y=1|X=0)}\right) = \beta$ u $OR = \exp(\beta)$. Por tanto probar la hipótesis H_0:OR=1 es equivalente a la hipótesis H_0:β=0.

De manera similar se obtiene en el caso de dos o más variables independientes la siguiente relación:

$$\frac{\Pr(Y=1|X')\Pr(Y=0|X)}{\Pr(Y=0|X')\Pr(Y=1|X)} = \exp\left(\sum_{i=1}^{p}\beta_i(X_i' - X_i)\right),$$

donde X=(X_1, X_2, ...,X_p).

Si el valor de X_i'=X_i, entonces el término $\beta_i(X_i' - X_i)$ es igual a cero y por tanto la expresión anterior no depende de X_i. Entonces si una de las variables, X_1 por ejemplo, representa la exposición a un factor de especial interés, el *OR* para individuos que son iguales en las restantes variables es OR=exp($\beta_1(X_1'$-$X_1)$), en particular si la variable X_1 está codificada como 1 si el factor está presente y como 0 si está ausente, entonces OR=exp(β_1). El odds ratio calculado de esta manera recibe el nombre de **odds ratio ajustado** por las variables X_2, ..., X_p.

Veamos el siguiente ejemplo del cálculo de *OR* ajustado.

EJEMPLO: Estudio de casos-controles de cáncer de pulmón y consumo de alcohol.

	Casos	Controles
Alcohol	68	32
No Alcohol	32	68

El odds ratio estimado es OR=4.52 con un intervalo de confianza igual a (2.39, 8.55). Si estratificamos por la variable fumar, obtenemos:

En Fumadores: *OR*=1.00 (0.21, 3.72)

	Casos	Controles
Alcohol	64	16
No Alcohol	16	4

En No fumadores: *OR*=1.00 (0.21, 3.72)

	Casos	Controles
Alcohol	4	16
No Alcohol	16	64

La variable FUMAR es un factor de confusión de la asociación entre cáncer de pulmón y consumo de alcohol. El *OR* de Mantel-Haenszel (2 estratos) = 1.0 (0.36, 2.49).

Veamos el mismo análisis con un modelo de regresión logística. El fichero de datos para el ejemplo anterior utilizando el programa RELODI es:

```
2,agrupados,cáncer,alcohol,fumar
1,64,1,1
1,16,0,1
0,16,1,1
0,4,0,1
1,4,1,0
1,16,0,0
0,16,1,0
0,64,0,0
```

Salida Abreviada de RELODI

```
Número de casos para los cuales cáncer es igual a 1: 100
Tamaño total de la muestra 200

Coeficiente   S.E.         z-score
-0.7537
 1.5075       0.3031        4.972609 alcohol

                           Interv de conf (95%)
Coefic.             Odds Ratio    Lim. inf.     Lim. sup.     Variable
 1.5057             4.5156        2.4926        8.1805        alcohol
```

En este caso solo consideramos la variable alcohol, y por tanto obtenemos un resultado similar a la primera tabla.

Salida Abreviada de RELODI

```
Coeficiente   S.E.         z-score
-1.3862
 0.0000       0.4419        0.0000 alcohol
 2.7725       0.4419        6.2735 fumar

                           Interv de conf (95%)
Coefic.             Odds Ratio   Lim. inf.     Lim. sup.     Variable
 0.0000             1.0000        0.4205        2.3778              alcohol
 2.7725             15.9991       6.7284        38.0436             fumar
```

Se obtiene entonces el *OR*=1.00 ajustado por la variable fumar, que es igual al *OR* de Mantel-Haenszel.

Si bien en un ejemplo como este en que sólo hay 2 variables independientes dicotómicas, el análisis estratificado es recomendable por su facilidad y comprensión, a medida que el número de variables crece o se consideran variables con más categorías, el análisis estratificado se hace muy laborioso. Por ejemplo, si consideramos 5 variables dicotómicas habría que calcular 2^4=16 tablas de 2x2. Si alguna de las variables independientes es continua se deberá clasificar la misma con la consiguiente pérdida de información, en esos casos la regresión logística es un procedimiento sumamente útil.

<u>Evaluación de la Interacción</u>:

Consideremos dos factores de exposición X_1 y X_2 (variables dicotómicas) podemos definir el riesgo $R_{ij} = \Pr(D = 1 | X_1 = i, X_2 = j)$ para los distintos niveles de exposición a X_1 y X_2, y calcular el *OR* para cada uno de estos niveles por: $OR_{ij} = \dfrac{R_{ij}.(1 - R_{00})}{R_{00}.(1 - R_{ij})}$.

La hipótesis nula de no interacción bajo un modelo multiplicativo es: H_0:OR_{11}=OR_{10} OR_{01}, que puede contrastarse utilizando el siguiente modelo de regresión logística:

$$\Pr(Y = 1 | X_1, X_2, X_1X_2) = \frac{1}{1 + \exp(-\alpha - \beta_1 X_1 - \beta_2 X_2 - \beta_3 X_1 X_2)},$$

pues se tiene la siguiente igualdad: $\beta_3 = \text{logit}\left(\dfrac{OR_{11}}{OR_{10}OR_{01}}\right)$.

EJEMPLO: Consideremos el siguiente estudio de cáncer de pulmón y los siguientes factores de exposición: X_1 FUMAR y X_2 VIVIR EN ZONA RURAL

Zona de Residencia	Fumadores		No fumadores	
	Casos	Controles	Casos	Controles
Rural	520	180	300	100
Urbana	30	220	150	500

El fichero de datos para este ejemplo utilizando el programa RELODI es:

```
3,agrupados,CANCER,FUMAR,CAMPO,CAMPO*FUMAR
1,520,1,1,1
1,30,0,1,0
1,300,1,0,0
1,150,0,0,0
0,180,1,1,1
0,220,0,1,0
0,100,1,0,0
0,500,0,0,0
```

Veamos los resultados de los siguientes modelos:

1) $\text{logit}(\Pr(Y = 1 | X_1)) = \alpha + \beta_1 X_1$, o sea considerando solo la variable FUMAR.

Salida Abreviada de RELODI

```
Coeficiente  S.E.         z-score
-1.3862
 2.4607      0.1083       22.7157 FUMAR

                   Interv de conf (95%)
Coefic.      Odds Ratio   Lim. inf.    Lim. sup.    Variable
 2.4607      11.7141      9.4732       14.4851      FUMAR
```

Como esperamos la variable fumar aparece asociada al cáncer de pulmón.

2) $\text{logit}(\Pr(Y=1|X_2)) = \alpha + \beta_2 X_2$, o sea considerando solo la variable VIVIR EN ZONA RURAL.

Salida Abreviada de RELODI

```
Coeficiente  S.E.         z-score
-0.2876
 0.6061      0.0905       6.6907 CAMPO

                   Interv de conf (95%)
Coefic.      Odds Ratio   Lim. inf.    Lim. sup.    Variable
 0.6061      1.8333       1.5350       2.1895       CAMPO
```

Algo que no esperamos, la variable vivir en zona rural aparece asociada al cáncer de pulmón. ¿Veamos si la variable FUMAR es de confusión?

Salida Abreviada de RELODI

```
Coeficiente  S.E.         z-score
-1.3104
 2.5751      0.1195       21.5457 FUMAR
-0.2912      0.1175       -2.4786 CAMPO

                   Interv de conf (95%)
Coefic.      Odds Ratio   Lim. inf.    Lim. sup.    Variable
 2.5751      13.1337      10.3908      16.6008      FUMAR
-0.2912      0.7473       0.5935       0.9408       CAMPO
```

Notemos que el *OR* ajustado por la variable FUMAR (*OR*=0.7473) nos indica que vivir en zona rural es un factor "protector" del cáncer de pulmón. FUMAR actúa como variable de confusión en esa relación.

3) $\text{logit}(\Pr(Y=1|X_1, X_2, X_1X_2)) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$.

Salida Abreviada de RELODI

```
Coeficiente  S.E.         z-score
-1.2039
 2.3025      0.1483       15.5240 FUMAR
-0.7884      0.2157       -3.6545 CAMPO
 0.7507      0.2595        2.8925 CAMPO*FUMAR

                   Interv de conf (95%)
Coefic.      Odds Ratio   Lim. inf.    Lim. sup.    Variable
 2.3025      10.0000      7.4772       13.3738      FUMAR
-0.7884      0.4545       0.2978       0.6937       CAMPO
 0.7507      2.1185       1.2738       3.5233       CAMPO*FUMAR
```

Notemos que el coeficiente β_3 es distinto de cero, por tanto se concluye que existe interacción entre ambos factores.

Si utilizamos el siguiente fichero de datos podemos estimar: $OR_{11} = \frac{R_{11}(1 - R_{11})}{R_{00}(1 - R_{11})}$:

```
1,agrupados,CANCER,CAMPO*FUMAR
1,520,1
1,150,0
0,180,1
0,500,0
```

4) $\text{logit}(\Pr(Y = 1 | X_1 X_2)) = \alpha + \beta_3 X_1 X_2$

Salida Abreviada de RELODI

```
Coeficiente  S.E.        z-score
-1.2039
 2.2648      0.1270       17.8243 CAMPO*FUMAR

                  Interv de conf (95%)
Coefic.      Odds Ratio   Lim. inf.          Lim. sup.    Variable
 2.2648      9.6295       7.5066             12.3528      CAMPO*FUMAR
```

Notemos que OR_{11}=9.6295 difiere de $OR_{01}OR_{10}$=1.8333*11.7141≈21.47, que sería el valor de OR_{11} si no hubiese interacción.

Ejercicio:

1.- Considere los siguientes resultados de estudio de cohortes donde se evalúa la exposición a dos factores E y F como posibles factores de riesgo de una enfermedad que denotaremos D.

	Expuestos a E		No expuestos a E	
	Casos	Controles	Casos	Controles
Expuestos a F	110	390	380	2620
No expuestos a F	90	1410	20	980

a) Mediante un modelo de regresión logística estime el *OR* crudo para los factores E y F.

b) Estime el *OR* ajustado de F controlando E y el *OR* ajustado de E controlando F.

c) ¿Alguno de los factores es de confusión?

d) ¿Existe interacción entre E y F?