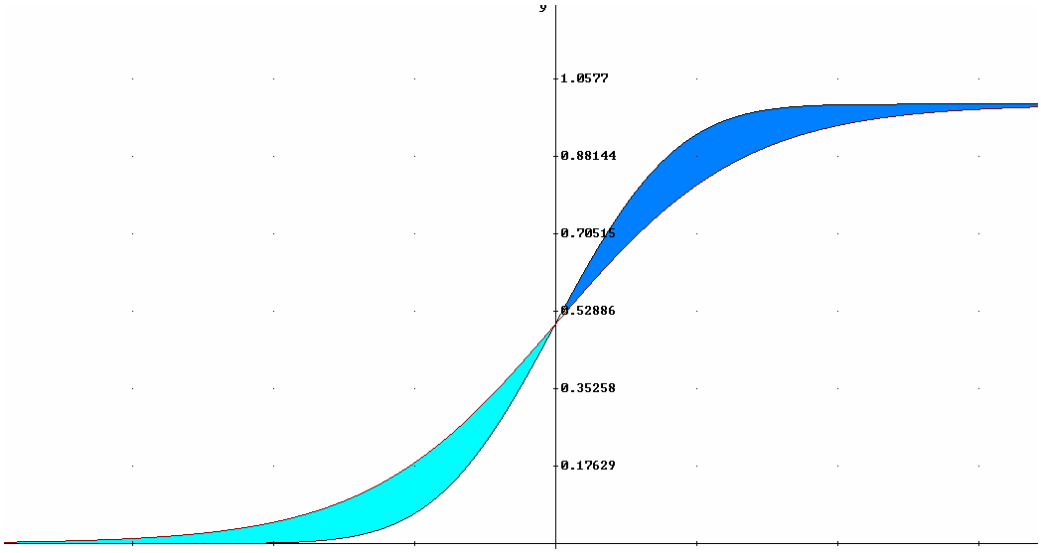
Regresión con variable dependiente cualitativa



J. M. Rojo Abuín
Instituto de Economía y Geografía
Madrid, II-2007

Índice

I. INTRODUCCIÓN

En muchas ocasiones estaremos interesados en predecir los valores de una variable dicotómica binaria, es decir, una variable que sólo puede tomar dos valores, los valores son complementarios y dichos valores no son comparables, como sucede en regresión lineal.

Ejemplos de variable dependiente dicotómica pueden ser: sano o enfermo, paga o no paga, …, etc.

El modelo de regresión logística se utiliza cuando estamos interesados en pronosticar la **probabilidad** de que ocurra o no un suceso determinado. Por ejemplo, a la vista de un conjunto de pruebas médicas, que una persona tenga una determinada enfermedad, o bien que un cliente devuelva un crédito bancario.

A diferencia del análisis discriminante que requiere la normalidad multivariante de los datos, el análisis de regresión logística sólo precisa del principio de **monotonía**, es decir, si el suceso A es que una determinada persona padezca de artrosis y X representa la edad, deberá de ocurrir:

$$x_i \geq x_j \Rightarrow P(A/x_i) \geq P(A/x_j)$$

A diferencia del análisis discriminante, podremos estudiar el **impacto** que tiene cada una de las variables explicativas en la probabilidad de que ocurra el suceso en estudio.

El análisis de regresión logística es una herramienta muy flexible en cuanto a la naturaleza de las variables explicativas, pues éstas pueden ser de escala y categóricas.

II. PLANTEAMIENTO DEL PROBLEMA

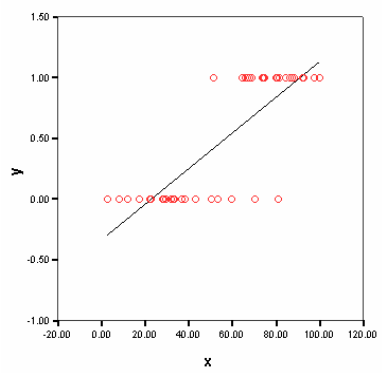
Supongamos que tenemos la variable de estudio codificada de la siguiente manera:

$$y = \begin{cases} 0 \text{ No ocurre el suceso} \\ 1 \text{ Si ocurre el suceso} \end{cases}$$

Además, vamos a considerar que sólo tenemos una variable explicativa X ; en estas condiciones podríamos considerar un modelo de regresión lineal con el propósito de ver qué dificultades nos van a surgir:

$$y_i = p_i = a + b * x_i + u_i$$

Si estimamos este modelo y representamos gráficamente la recta de regresión:



Podemos observar que la línea de regresión no está acotada en el intervalo [0,1] y, por lo tanto, ya no va a representar una probabilidad.

Además, consideraciones de índole matemática nos llevan a la conclusión de que los residuos no van a ser homocedásticos y, por tanto, la técnica de estimación por mínimos cuadrados dejará de ser un método óptimo de estimación.

Una forma que tenemos de garantizar que los valores pronosticados estén en el intervalo [0,1] es considerar la siguiente transformación:

$$p(a/x) = F(x * b)$$

Donde F es una función de distribución.

Nota

Una función de distribución es una función real de variable real:

$F : R \to R$

De forma que verifica:

EstÁ acotada en el intervalo [0,1]

$0 \le F(x) \le 1 \ \forall \ x$

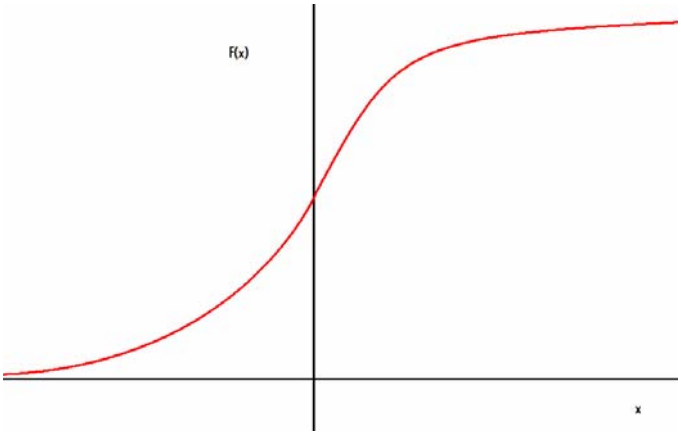
Es monótona no decreciente:

$x_1 \le x_2 \ \Rightarrow \ F(x_1) \le F(x_2)$

Y, además, está definida en todo R, tomando los siguientes valores:

$F(-\infty) = 0$
$F(+\infty) = 1$

En general, la grafica de una función de distribución es:



Si utilizamos la función de distribución logística, el análisis se denomina **Regresión Logística**, y si utilizamos la función de distribución normal se denomina **Regresión Probit**.

III. EL MODELO DE REGRESIÓN LOGÍSTICA

El modelo de regresión logística parte de la hipótesis de que los datos siguen el siguiente modelo:

$$\ln(\frac{p}{1-p}) = b_0 + b_1 * x_1 + b_2 * x_2 + ... + b_k * x_k + u = x * b + u$$

Con el fin de simplificar la notación, definimos Z:

$$z = b_0 + b_1 * x_1 + b_2 * x_2 + ... + b_k * x_k$$

Por lo tanto, el modelo se puede representar como:

$$\ln(\frac{p}{1-p}) = z + u$$

Donde **p** es la probabilidad de que ocurra el suceso de estudio.

Operando algebráicamente sobre el modelo:

$$\ln(\frac{p}{1-p}) = z$$

$$\frac{p}{1-p} = e^z$$

$$p = (1-p) * e^z$$

$$p = e^z - p * e^z$$

$$p + p * e^z = e^z$$

$$p(1+e^z) = e^z$$

$$p = \frac{e^z}{1+e^z}$$

Como la función de distribución logística es:

$$F(x) = \frac{e^x}{1+e^x}$$

Por tanto, podemos reescribir el modelo de forma mucho más compacta:

$$p = \frac{e^z}{1+e^z} = F(z) = F(x*b)$$

De donde se deduce que el modelo de regresión logística es, en principio, un modelo de **regresión no lineal**, pero es lineal en escala logarítmica atendiendo a su definición original:

$$\ln(\frac{p}{1-p}) = z$$

$$\ln(p) - \ln(1-p) = z$$

$$\ln(p) - \ln(1-p) = b_0 + b_1 * x_1 + b_2 * x_2 + ... + b_k * x_k$$

Es decir, la diferencia de la probabilidad de que ocurra un suceso respecto de que no ocurra es lineal pero en escala logarítmica. Por tanto, el significado de los coeficientes, aunque guardando una cierta relación con el modelo de regresión lineal, va a ser algo más complejo de interpretar.

Recordemos las dos formas más importantes de expresar el modelo de regresión logística:

$$\ln(p) - \ln(1-p) = b_0 + b_1 * x_1 + b_2 * x_2 + ... + b_k * x_k$$

$$\frac{p}{1-p} = e^{b_0} * e^{b1*X_1} * e^{b2*X_2} ... e^{bk*X_k}$$

La primera expresión se llama **logit** y a la segunda **Odds ratio** o cociente de probabilidades.

IV. ESTIMACIÓN DE LOS PARÁMETROS.

Brevemente, vamos a ver en esquema el problema que ofrece, en el caso de regresión logística, la estimación de los parámetros.

Sea una muestra de **n** elementos, donde se ha observado la variable respuesta Y (que sólo puede tomar dos valores: cero y uno) y la variable **X** .

La función de probabilidad de una observación cualquiera es:

$$P(Y=1/x)=p$$

$$P(Y=0/x)=1-p$$

Por tanto:

$$P(Y/x)=p^{y}*(1-p)^{1-y}$$

Por tanto la función de probabilidades de la muestra es:

$$P(y_1,y_2,...,y_n)=\prod_i p_i^{Y_i}*(1-p_i)^{1-y_i}$$

Esta expresión recibe el nombre de verosimilitud de la muestra (likelihood).

Tomando logaritmos:

$$\log P(Y)=\sum_i^n y_i\cdot Log(\frac{p_i}{1-p_i})+\sum_i^n \log(1-p_i)$$

Expresando p_i en función de los parámetros que deseamos estimar:

$$L(B)=\sum_i^n y_i*x_i*b-\sum Log(1+e^{x_i*b})$$

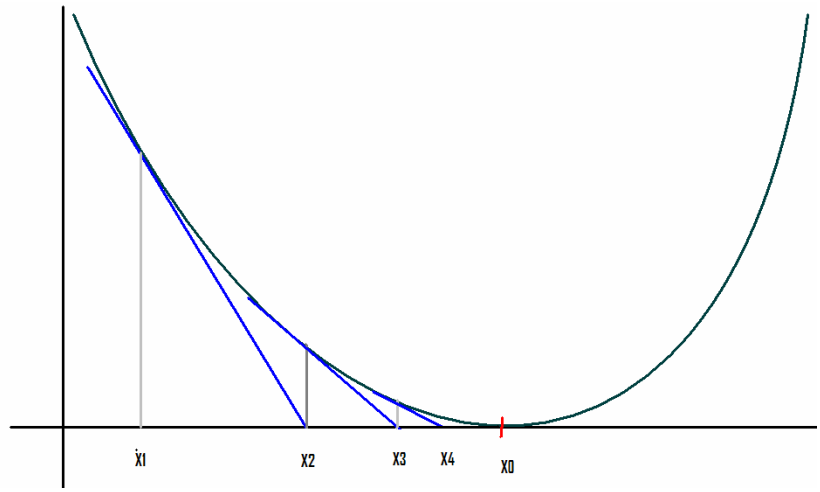
Resulta obvio que aunque derivemos y establezcamos la condición de máximo, no vamos a poder despejar los coeficientes **B** .

La solución que vamos a obtener es:

$$B_a = B_0 + \left(-\frac{\partial^2 L(B)}{\partial B * \partial B'} \right)^{-1} * \left(\frac{\partial L(B)}{\partial B} \right)$$

Esta solución establece cómo encontrar una solución (B_a) a partir de un punto próximo cualquiera, denominado B_0 . Por lo tanto, deberemos de hacer una estimación inicial del valor de los verdaderos parámetros y mediante un procedimiento recursivo encontrar el verdadero valor de los mismos. Para encontrar los verdaderos valores se suele utilizar el algoritmo de **Newton-Raphson**.

Gráficamente:



V. EJEMPLO 1

Vamos a ir introduciendo los elementos de esta técnica a través de un sencillo ejemplo.

El tratamiento y pronóstico del cáncer depende de cuánto se haya extendido la enfermedad.

Unas de las zonas propensas a ser afectadas por la enfermedad son los ganglios linfáticos.

Si los ganglios linfáticos están afectados el tratamiento pierde efectividad.

Para ciertos tipos de cáncer es preciso realizar una intervención quirúrgica para determinar si la enfermedad se ha extendido al sistema linfático, y así determinar qué tratamiento se deberá de aplicar.

Si en función a una serie de pruebas médicas no invasivas se pudiera determinar si los ganglios linfáticos están afectados o no se ahorraría tiempo y molestias a los pacientes.

Los datos que vamos a analizar pertenecen a una muestra aleatoria de 53 pacientes masculinos con cáncer de próstata. A cada paciente se le han medido las siguientes variables o características:

- Xray: Resultado de la prueba de rayos X
- Grado: Grado de agresividad del tumor.
- Estado: Cómo está de extendida la enfermedad.
- Nodos: Indicador de si los ganglios linfáticos están afectados o no por la enfermedad.
- Edad: edad del paciente.
- Acido: Prueba de laboratorio del nivel de ácido phosphatase.

A continuación mostramos los estadísticos descriptivos de las variables involucradas en el análisis. Es de particular importancia asegurarse que las variables del tipo ausencia/presencia estén codificadas como cero y uno.

xray Prueba de rayos X

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 Negativo	38	71,7	71,7	71,7
	1 Positivo	15	28,3	28,3	100,0
	Total	53	100,0	100,0	

grado Grado de agresividad

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 No agresivo	33	62,3	62,3	62,3
	1 Agresivo	20	37,7	37,7	100,0
	Total	53	100,0	100,0	

estado Estado de la enfermedad

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 No extendido	26	49,1	49,1	49,1
	1 Extendido	27	50,9	50,9	100,0
	Total	53	100,0	100,0	

nodos Estado de los ganglios linfaticos

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 No afectados	33	62,3	62,3	62,3
	1 Afectados	20	37,7	37,7	100,0
	Total	53	100,0	100,0	

Statistics

		edad Edad en años	acid Acido phosphatase
N	Valid	53	53
	Missing	0	0
Mean		59,38	69,42
Median		60,00	65,00
Mode		56[a]	50
Std. Deviation		6,168	26,201
Skewness		-,495	2,252
Std. Error of Skewness		,327	,327
Kurtosis		-,697	7,295
Std. Error of Kurtosis		,644	,644
Minimum		45	40
Maximum		68	187

a. Multiple modes exist. The smallest value is shown

V.1. Coeficientes estimados del modelo logístico

Con las variables anteriores vamos a intentar construir un modelo de regresión logística para tratar de pronosticar en qué pacientes se encuentran los ganglios linfáticos (nodos) afectados por la enfermedad.

Coeficientes del modelo de regresión logística.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	edad	-,069	,058	1,432	1	,231	,933
	acid	,024	,013	3,423	1	,064	1,025
	xray	2,045	,807	6,421	1	,011	7,732
	grado	,761	,771	,976	1	,323	2,141
	estado	1,564	,774	4,084	1	,043	4,778
	Constant	,062	3,460	,000	1	,986	1,064

a. Variable(s) entered on step 1: edad, acid, xray, grado, estado.

En la **segunda columna** se muestran los coeficientes estimados **B**. Para poder interpretar dichos coeficientes hay que tener en cuenta como están codificadas las variables, pues las dos primeras: edad y acido son contínuas y el resto están codificadas como 0 o 1, para indicar ausencia o presencia de una determinada característica.

En **la tercera columna** es muestra la desviación típica del estimador.

La **cuarta columna** muestra el **estadístico de Wald**; el estadístico de Wald es:

$$W(b_j) = \left(\frac{\hat{b}_j}{\sigma(b_j)} \right)^2$$

y dicho estadístico se distribuye de acuerdo con una χ_1^2 ; por tanto, todos los coeficientes que tengan un $W(b_j) > 4$ serán significativos.

La **sexta columna** (sig) es el **p-value** del coeficiente.

La **séptima columna** es el **exponencial del coeficiente**. El interés del exponencial de los coeficientes es el estudio del impacto de las variables cualitativas.

Ejemplo.

En este ejemplo hemos codificado la prueba de **rayos x** de forma dicotómica (0,1). Por tanto:

$$\frac{p}{1-p} = e^{z} = k * e^{2.045 * xray}$$

Si la prueba de rayos X es negativa, la variable vale 0, y si es positiva la variable vale 1; por tanto, si la prueba de rayos x es positiva el cociente de probabilidades aumenta:

$$e^{2.045} = 7.73$$

Pues:

$$\frac{p}{1-p} = e^{z} = k * e^{2.045 * xray}$$

Resultado negativo de la prueba:

$$\frac{p}{1-p} = e^{z} = k * e^{2.045 * 0} = k * 1$$

Resultado positivo de la prueba:

$$\frac{p}{1-p} = e^{z} = k * e^{2.045 * 1} = k * 7.73$$

Luego, si la prueba de rayos x es positiva, la probabilidad de tener el sistema linfatico afectado queda multiplicada por 7.73.

V.2. Estimando probabilidades

Con los coeficientes estimados ya es posible predecir la probabilidad de que una persona tenga los ganglios linfáticos afectados por el cáncer simplemente construyendo la función de probabilidad:

$$P(nodo = 1/x) = \frac{1}{1+e^{-z}}$$

Donde:

$$Z = 0.62 + 1.56 * estado + 0.76 * grado + 2 * xray + 0.024 * acido - 0.07 * edad$$

A la vista de esta ecuación, podemos estimar la probabilidad de que un hombre con determinadas características tenga el sistema linfático afectado.

Por ejemplo, la probabilidad de que un hombre de 60 años, con un nivel de ácido de 50 y negativo en el resto de las pruebas es de:

$$Z = 0.62 + 1.56 * estado + 0.76 * grado + 2 * xray + 0.024 * acido - 0.07 * edad$$

$$Z = 0.62 + 1.56 * 0 + 0.76 * 0 + 2 * 0 + 0.024 * 50 - 0.07 * 60$$

$$Z = -2.38$$

$$P(nodo = 1/x) = \frac{1}{1+e^{-(-2.38)}} = 0.085$$

En cambio la misma persona dando positivo en todas las pruebas va a tener una probabilidad estimada de:

$$Z = 0.62 + 1.56 * 1 + 0.76 * 1 + 2 * 1 + 0.024 * 50 - 0.07 * 60$$

$$Z = 1.94$$

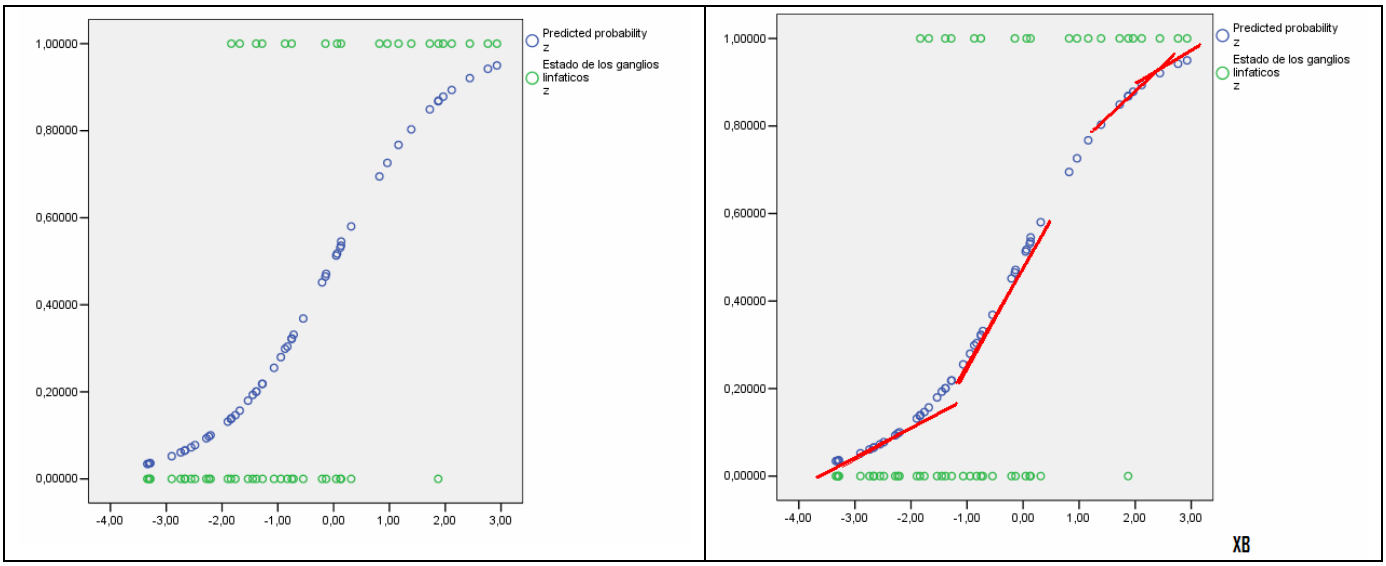
$$P(nodo = 1/x) = \frac{1}{1+e^{-(1.94)}} = 0.87$$

V.3. Interpretando los coeficientes

Si bien en regresión lineal la interpretación de los coeficientes de regresión es simple e intuitiva:

B_k es el incremento producido en la variable dependiente por un incremento unitario en la variable X_k.

En la regresión logística no va a ser tan intuitiva, al depender tanto del valor de X_k donde se produzca el incremento como del valor del resto de las variables, pues la pendiente de la curva de regresión va a ir variando.



Para ayudar a interpretar los coeficientes de regresión logística definimos el **Odds Ratio** como el cociente de probabilidades entre que ocurra un suceso respecto de que no ocurra:

$$OddRatio = \frac{P(Y=1)}{P(Y=0)} = \frac{P}{1-P}$$

Teniendo en cuenta que el modelo de regresión logística puede ser escrito como:

$$\ln(p) - \ln(1-p) = b_0 + b_1 * x_1 + b_2 * x_2 + ... + b_k * x_k$$

$$\ln(\frac{p}{1-p}) = b_0 + b_1 * x_1 + b_2 * x_2 + ... + b_k * x_k$$

Los coeficientes **B** indican el incremento de la probabilidad de que ocurra el suceso, es decir, la probabilidad de que el sistema linfático esté afectado respecto de que no esté afectado pero en escala logarítmica.

Si el coeficiente p-esimo vale cero, indica que la variable p-esima no afecta a la ocurrencia del suceso.

Si el coeficiente p-esimo es negativo indica que a media que dicha variable va aumentando va a ir disminuyendo el logaritmo del cociente de probabilidades y al revés si es positivo.

Si tomamos exponenciales:

$$\frac{p}{1-p} = e^{B_0} * e^{B_1 * x_1} * e^{B_2 * x_2} * ... * e^{B_k * x_k}$$

Por tanto el coeficiente e^{B_p} va a significar por cuánto se multiplica el Odds Ratio.

Otra forma de verlo algo más intuitiva es considerar la derivada de la función de regresión respecto de la p-esima variable.

Tenemos que la probabilidad de ocurrencia del evento es una función de X y B:

$$P(Y = 1/X) = \Lambda(X, B) = \Lambda(X * B)$$

Si derivamos respecto de la p-esima variable:

$$\frac{\partial \Lambda}{\partial x_p} = \Lambda'(X * B)b_p = \Lambda'(b_0 + b_1 x_1 + ...b_k x_k)b_p$$

El problema es que la derivada va a depender de qué valor tomamos para las k variables, es decir, en qué punto vamos a evaluar la curva.

Podemos evaluarla en el punto medio:

$$\boxed{\Lambda'(b_0 + b_1 \bar{x}_1 + ...b_k \bar{x}_k)b_p}$$

O bien podemos considerar la media de los incrementos:

$$\boxed{\sum_i \frac{\Lambda'(b_0 + b_1 x_{i,1} + ...b_k x_{i,k})b_p}{n}}$$

El significado de la primera expresión es el incremento en la probabilidad de ocurrencia del suceso en una persona **media** por un incremento unitario en la p-esima variable.

La segunda expresión indica cuál es la **media de los incrementos** de la probabilidad de ocurrencia del suceso por un incremento unitario en la p-esima variable.

Las dos últimas formas no están implementadas en la aplicación SPSS y deberemos realizarlas a mano. En el caso que nos ocupa:

Media de los incrementos:

Report

	edadB	acidB	xrayB	gradoB	estadoB
Mean	-,0101	,0035	,2994	,1114	,2284
N	53	53	53	53	53
Std. Deviation	,00504	,00175	,14944	,05561	,11400

Código en SPSS

```
compute z= -0.069*edad+0.024*acid+2.045*xray+0.761*grado+1.564*estado+0.062.
execute.
compute edadB=      1/(1+exp(-z))**2*exp(-z)*(-0.069).
compute acidB=      1/(1+exp(-z))**2*exp(-z)*(0.024).
compute xrayB=      1/(1+exp(-z))**2*exp(-z)*(2.045).
compute gradoB=     1/(1+exp(-z))**2*exp(-z)*(0.761).
compute estadoB=    1/(1+exp(-z))**2*exp(-z)*(1.56).
execute.
mean edadB to estadoB.
```

Incremento en una persona media:

Report

	edadB	acidB	xrayB	gradoB	estadoB
Mean	-,0152	,0053	,4517	,1681	,3446
N	53	53	53	53	53
Std. Deviation	,00000	,00000	,00000	,00000	,00000

Código en SPSS

```
compute z= -0.069*59.4+0.024*69.42+2.045*0.28+0.761*0.38+1.564*0.51+0.062.
compute edadB=      1/(1+exp(-z))**2*exp(-z)*(-0.069).
compute acidB=      1/(1+exp(-z))**2*exp(-z)*(0.024).
compute xrayB=      1/(1+exp(-z))**2*exp(-z)*(2.045).
compute gradoB=     1/(1+exp(-z))**2*exp(-z)*(0.761).
compute estadoB=    1/(1+exp(-z))**2*exp(-z)*(1.56).
execute.
mean edadB to estadoB.
```

VI. EJEMPLO COMPLETO

Seguimos con el ejemplo anterior, pero mostrando tanto estadísticos de bondad de ajuste como los de contraste de regresión.

En primer lugar se muestran los esquemas de codificación de las variables, tanto la variable respuesta como las variables categóricas:

Dependent Variable Encoding

Original Value	Internal Value
0 No afectados	0
1 Afectados	1

A la vista del esquema de codificación de la variable respuesta, el modelo va a tratar de predecir la probabilidad de que una persona tenga el sistema linfático afectado.

Categorical Variables Codings

		Frequency	Paramete (1)
estado Estado de la enfermedad	0 No extendido	26	,000
	1 Extendido	27	1,000
grado Grado de agresividad	0 No agresivo	33	,000
	1 Agresivo	20	1,000
xray Prueba de rayos X	0 Negativo	38	,000
	1 Positivo	15	1,000

En el resto de las variables categóricas vemos que el esquema de codificación interno coincide con el externo.

VI.1. Historial de las iteracciones

Iteration History[a,b,c,d,e]

Iteration		-2 Log likelihood	Coefficients		
			Constant	xray(1)	estado(1)
Step 1	1	59,116	-1,053	1,986	
	2	59,001	-1,167	2,177	
	3	59,001	-1,170	2,182	
	4	59,001	-1,170	2,182	
Step 2	1	54,101	-1,564	1,735	1,144
	2	53,366	-1,979	2,069	1,527
	3	53,353	-2,043	2,118	1,587
	4	53,353	-2,045	2,119	1,588
	5	53,353	-2,045	2,119	1,588

a. Method: Forward Stepwise (Conditional)

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 70,252

d. Estimation terminated at iteration number 4 because parameter estimates changed by less than ,001.

e. Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

Realiza dos pasos, por lo tanto se han introducido dos variables en el modelo.

En cada paso va aumentando la verosimilitud del modelo, lo cual implica que disminuye la siguiente expresión: -2 log(verosimilitud) (-2LL).

En los dos pasos el algoritmo termina correctamente porque se alcanza el criterio de parada, es decir, el cambio entre los coeficientes estimados en a ultima iteración es inferior a 0.001.

VI.2. Contraste de regresión

El contraste de regresión en estos modelos no se realiza sobre la descomposición de la suma de cuadrados como en regresión lineal sino sobre el incremento de la verosimilitud, mas exactamente sobre la disminución de -2LL.

Hipótesis

$$H_0 \;\; b_1 = b_2 = ... = b_k = 0$$

$$H_1 \;\; \exists \;\; b_p \neq 0$$

Construcción del contraste

$$C2LL = -2LL(b_0) - (-2LL(b_0, b_1, b_2, ..., b_k)$$

La diferencia de verosimilitudes se distribuye de acuerdo con una distribución χ_j^2, donde J es la diferencia del número de parámetros en el modelo.

Iteration History[a,b,c]

Iteration		-2 Log likelihood	Coefficients Constant
Step 0	1	70,253	-,491
	2	70,252	-,501
	3	70,252	-,501

a. Constant is included in the model.

b. Initial -2 Log Likelihood: 70,252

c. Estimation terminated at iteration number 3 because parameter estimates changed by less than ,001.

La verosimilitud con sólo la constante es de 70.252.

Iteration History[a,b,c,d,e]

Iteration		-2 Log likelihood	Coefficients Constant	xray(1)	estado(1)
Step 1	1	59,116	-1,053	1,986	
	2	59,001	-1,167	2,177	
	3	59,001	-1,170	2,182	
	4	59,001	-1,170	2,182	
Step 2	1	54,101	-1,564	1,735	1,144
	2	53,366	-1,979	2,069	1,527
	3	53,353	-2,043	2,118	1,587
	4	53,353	-2,045	2,119	1,588
	5	53,353	-2,045	2,119	1,588

a. Method: Forward Stepwise (Conditional)

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 70,252

d. Estimation terminated at iteration number 4 because parameter estimates changed by less than ,001.

e. Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

La verosimilitud (-2LL) con una sola variable es de 59.001.

La verosimilitud (-2LL) con dos variables es de 53.353.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	11,251	1	,001
	Block	11,251	1	,001
	Model	11,251	1	,001
Step 2	Step	5,647	1	,017
	Block	16,899	2	,000
	Model	16,899	2	,000

La primera variable que entra produce una disminución en -2LL de:

70.252-59.001= 11.251.

P($X^2 > 11.252$) =0.00079

Por lo tanto rechazamos la hipótesis nula y aceptamos que la primera variable es significativa.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	11,251	1	,001
	Block	11,251	1	,001
	Model	11,251	1	,001
Step 2	Step	5,647	1	,017
	Block	16,899	2	,000
	Model	16,899	2	,000

La segunda variable sobre la primera produce una reducción de -2LL de:

59.001-53.353=5.647

P($\chi_1^2 > 5.647$) = 0.01747

Por lo tanto la introducción de la segunda variable sigue siendo significativa.

VI.3. Medidas de bondad del ajuste

En este tipo de modelos no se emplea el R^2 para mostrar la bondad del ajuste, sino que se calcula el incremento de la verosimilitud, aunque reciben el nombre de R^2 no van a tener el significado geométrico que tienen en regresión lineal por lo tanto deberían de llamarse pseudos R^2.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	59,001[a]	,191	,260
2	53,353[b]	,273	,372

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than ,001.

b. Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

Cox y Snell:

$$R^2 = 1 - \frac{L(b_0)}{L(b_0, b_1, ...b_k)}$$

$$R^2 = 1 - \left(\frac{L(b_0)}{L(b_0, b_1, ...b_k)} \right)^{\frac{2}{N}}$$

En este ejemplo:

$$R^2 = 1 - (\frac{70.25}{53.353})^{\frac{2}{53}} = 0.273$$

Este coeficiente está acotado:

$$0 \leq R^2 < 1$$

Es decir, no puede alcanzar el valor 1.

R cuadrado de Nagelkerke

Nagelkerke prefiere definir R^2 como:

$$R^2 = \frac{R^2}{R^2_{Max}}$$

Donde $R^2_{Max} = 1 - \left(L(b_0)\right)^{\frac{2}{N}}$

Para así poder alcanzar el valor 1.

Aunque estos coeficientes tratan de medir la variabilidad explicada, en general, van a ser mucho más bajos que en regresión lineal y deberán de ser complementados con otras medidas de bondad de ajuste.

Test de Hosmer y Lemeshow

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	,000	0	.
2	,798	2	,671

Contingency Table for Hosmer and Lemeshow Test

		nodos Estado de los ganglios linfaticos = 0 No afectados		nodos Estado de los ganglios linfaticos = 1 Afectados		
		Observed	Expected	Observed	Expected	Total
Step 1	1	29	29,000	9	9,000	38
	2	4	4,000	11	11,000	15
Step 2	1	18	18,593	3	2,407	21
	2	11	10,407	6	6,593	17
	3	3	2,407	2	2,593	5
	4	1	1,593	9	8,407	10

El test de Hosmer y Lemeshow es un constaste de distribución.

La hipótesis nula es que no hay diferencias entre los valores observados y los valores pronosticados (probabilidades); la alternativa es que sí las hay. Por tanto, el rechazo de este test indica que el modelo no está bien ajustado.

En este caso la significatividad de este es de 0.798, no rechazamos la hipótesis nula y por tanto no rechazamos que el modelo tiene falta de ajuste.

Tabla de clasificación

Si bien los coeficientes de bondad de ajuste no son del todo fiables, la tabla de clasificación es normalmente el criterio que debemos de seguir para indicar la bondad de ajuste del modelo.

En esta tabla se muestran los casos bien clasificados en la diagonal principal, y los casos mal clasificados en la segunda diagonal.

Classification Table[a]

			Predicted		
			Estado de los ganglios linfaticos		Percentage Correct
	Observed		No afectados	Afectados	
Step 1	Estado de los ganglios linfaticos	No afectados	29	4	87,9
		Afectados	9	11	55,0
	Overall Percentage				75,5
Step 2	Estado de los ganglios linfaticos	No afectados	29	4	87,9
		Afectados	9	11	55,0
	Overall Percentage				75,5

a. The cut value is ,500

De las 29 + 4 personas que **no** tienen los ganglios afectados, 29 han sido pronosticados como sanos, es decir, un porcentaje de aciertos del $\frac{29}{33} = 87\%$

De las 9 + 11 personas que **sí** tienen los ganglios afectados, 11 han sido pronosticados como afectados, un porcentaje de aciertos del $\frac{11}{20} = 55\%$

El porcentaje global de aciertos es del $\frac{29+11}{29+11+4+9} = 75.5\%$