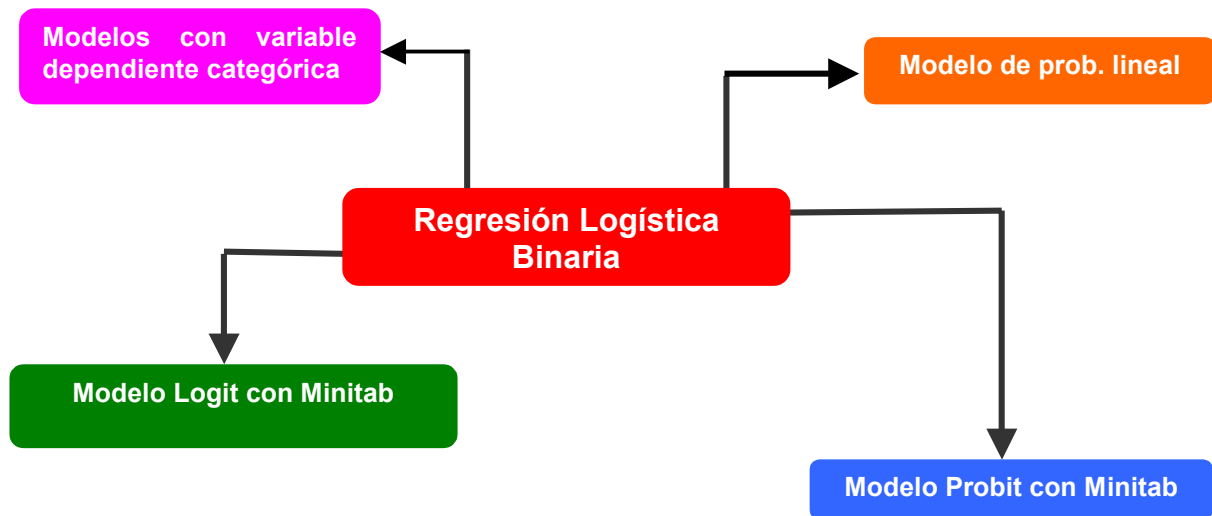


# REGRESIÓN LOGÍSTICA BINARIA

**Autores:** Ángel Alejandro Juan Pérez (ajuanp@uoc.edu), Renatas Kizys (rkizys@uoc.edu), Luis María Manzanedo Del Hoyo (lmanzanedo@uoc.edu).

## ESQUEMA DE CONTENIDOS



## INTRODUCCIÓN

Una variable binaria es aquella que sólo puede adquirir dos posibles valores (Sí-No, 0-1, Verdadero-Falso, etc.). Las variables binarias constituyen un subconjunto muy importante de las llamadas variables categóricas o cualitativas, las cuales están muy presentes en la economía y las ciencias sociales. En concreto, este tipo de variables juegan un papel fundamental en áreas como la teoría de la decisión y el *management*.

Cuando se pretende explicar, mediante un modelo de regresión, el comportamiento de una variable (llamada variable endógena o dependiente) en función de los valores que tomen otras (llamadas variables exógenas o explicativas), suele utilizarse un modelo de regresión lineal múltiple (MRLM o MRLG). Ahora bien, como veremos en este *math-block*, el modelo lineal presenta ciertos problemas serios cuando la variable dependiente es binaria (y, en general, categórica), lo cual nos llevará a usar modelos de regresión no lineales -específicamente pensados para realizar regresión con variables categóricas. Los modelos que analizaremos aquí serán el Logit y el Probit.

## OBJETIVOS

---

- Ampliar los conceptos de regresión al caso en que la variable dependiente sea categórica.
- Conocer el modelo de probabilidad lineal y los problemas que éste presenta a la hora de explicar el comportamiento de una variable dependiente binaria.
- Entender los modelos Logit y Probit como modelos que permiten superar las dificultades del modelo de probabilidad lineal.
- Aprender a realizar regresión logística binaria con ayuda de Minitab, interpretando correctamente los resultados generados por el programa.

## CONOCIMIENTOS PREVIOS

---

Aparte de estar iniciado en el uso del paquete estadístico Minitab, resulta muy conveniente haber leído con profundidad los siguientes *math-blocks*:

- Regresión Lineal Múltiple
- Introducción al MRLG

## CONCEPTOS FUNDAMENTALES

---

### □ Modelos con variable dependiente cualitativa

Un modelo de regresión múltiple (no necesariamente lineal) nos permite explicar el comportamiento de una variable dependiente  $Y$  en función de una serie de variables independientes  $X_1, X_2, \dots, X_k$  y de un término de perturbación  $u$ , i.e.:

$$Y = f(X_1, X_2, \dots, X_k, u)$$

En el caso particular de que el modelo de regresión sea lineal, tendremos una expresión de la forma:

$$Y = \beta_1 + \beta_2 \cdot X_2 + \dots + \beta_k \cdot X_k + u$$

(donde usamos la notación  $X_1 = 1$  para la “variable” que acompaña al término independiente).

El objetivo de la regresión será estimar los parámetros del modelo (en el caso lineal:  $\beta_1, \beta_2, \dots, \beta_k$ ), de forma que el modelo resultante se ajuste lo mejor posible a las observaciones.

Cuando la variable dependiente  $Y$  es continua, resulta frecuente utilizar un modelo de regresión lineal múltiple como el anterior. En tal caso la estimación de los parámetros  $\beta_1, \beta_2, \dots, \beta_k$  se lleva a cabo mediante los métodos de Mínimos Cuadrados (MCO o MCG).

Por otro lado, puede ocurrir que la variable dependiente  $Y$  sea una variable cualitativa o categórica, i.e., que  $Y$  sólo pueda tomar un conjunto reducido de valores. En tales circunstancias, el modelo de regresión lineal presentará una serie de inconvenientes serios

(como veremos en este *math-block*), por lo que será necesario recurrir a la llamada regresión logística.

A diferencia de la regresión lineal (que, como hemos dicho, suele hacer uso de los métodos de estimación por Mínimos Cuadrados), en la regresión logística se emplean los métodos de Máxima Verosimilitud (MV) para llevar a cabo la estimación de los parámetros del modelo.

En economía, estos modelos de regresión con variable endógena categórica suelen emplearse para explicar la decisión  $Y$  que toma un individuo -de entre un número limitado de posibles opciones- a partir de un conjunto de variables explicativas  $X_1, X_2, \dots, X_k$ . Así, p.e., supongamos que unos grandes almacenes han de seleccionar una de entre cinco posibles ciudades para ubicar su nueva sede. A fin de optimizar su proceso de elección, sería posible construir un modelo de regresión lineal múltiple que permitiese seleccionar la eventual ubicación en función de una serie de variables explicativas como pueden ser: el tamaño de la población, el número de otros centros de características similares en la zona, la renta per cápita, etc. Por este motivo, los modelos de variable endógena cualitativa son también llamados **modelos de elección discreta**.

Dentro de las variables categóricas, podemos distinguir varios tipos:

- Variables categóricas binarias: son aquellas que sólo pueden tomar dos valores (Éxito-Fracaso, 0-1, Sí-No etc.)
- Variables categóricas ordinales: pueden tomar múltiples valores, entre los cuales es posible establecer una relación de orden (Ninguno-Alguno-Muchos, Primero-Segundo-Tercero-Cuarto, Pequeño-Mediano-Grande-Muy Grande, etc.)
- Variables categóricas nominales: pueden tomar múltiples valores, si bien no es posible ordenarlos (Azul-Rojo-Verde-Blanco, Madrid-Sevilla-Barcelona-Alicante-Bilbao, etc.)

En este *math-block* vamos a centrarnos en el análisis del primer tipo de variable categórica, el de variable binaria.

## □ El modelo de probabilidad lineal y sus problemas

Consideremos el caso de una variable dependiente **binaria**,  $Y$ , la cual viene explicada por un conjunto de predictores  $X_1, X_2, \dots, X_k$ .

Observar que, por ser  $Y$  una variable binaria (i.e.: sólo podrá tomar los valores 0 y 1), siempre se cumplirá que:

$$E[Y] = 0 \cdot P(Y = 0) + 1 \cdot P(Y = 1) = P(Y = 1)$$

Por otra parte, podemos pensar en utilizar un modelo de regresión lineal múltiple para explicar el comportamiento de la variable  $Y$ , i.e.:

$$Y = \beta_1 + \beta_2 X_2 + \dots + \beta_k X_k + u$$

Bajo el supuesto habitual de que  $E[u] = 0$ , y suponiendo conocidos los valores que toman las variables explicativas (observaciones), tendremos que:

$$E[Y] = \beta_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

Igualando las dos expresiones obtenidas para  $E[Y]$  llegamos al resultado que le da nombre al **modelo de probabilidad lineal**:

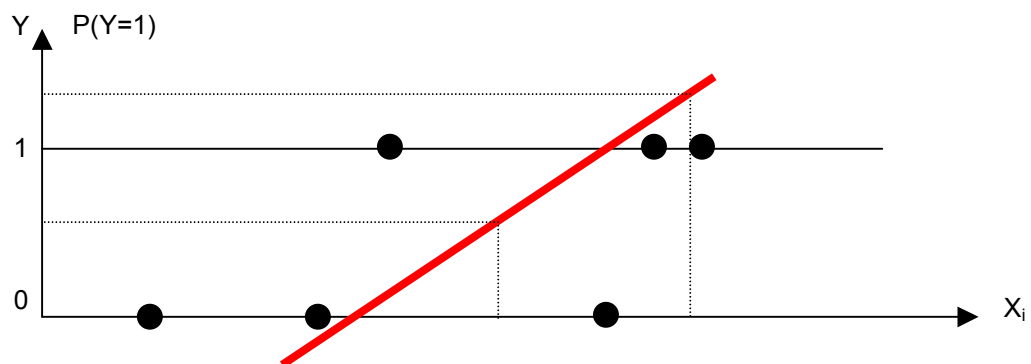
$$P(Y = 1) = \beta_1 + \beta_2 X_2 + \dots + \beta_k X_k = Y - u$$

Observar que esta expresión nos viene a decir que podemos expresar la variable dependiente **binaria**  $Y$  como la probabilidad de “éxito” más un término de perturbación, i.e.:

$$Y = P(Y = 1) + u = E[Y] + u$$

Sin embargo, este modelo inicial no será válido para explicar el comportamiento de variables dependientes binarias, pues presenta varios problemas:

1. El término de perturbación  $u = Y - (\beta_1 + \beta_2 X_2 + \dots + \beta_k X_k)$  ya no será una variable aleatoria continua (como ocurría en el MRLM), sino que será una variable aleatoria discreta –puesto que, conocidos los valores de las variables explicativas,  $u$  sólo puede tomar dos valores determinados. Por tanto,  $u$  ya no se distribuirá de forma normal (uno de los supuestos básicos del MRLM). Si bien este supuesto no resulta estrictamente necesario para aplicar MCO, sí es fundamental a la hora de realizar cualquier tipo de inferencia posterior sobre el modelo (intervalos de confianza para los parámetros estimados, contrastes de hipótesis, etc.).
2. El término de perturbación  $u$  no cumple la hipótesis de homoscedasticidad (i.e.: la varianza de dicho término no es constante). Debido a este problema, los estimadores MCO no serán eficientes, por lo que resultará necesario recurrir a la estimación por MCG.
3. Como la variable dependiente  $Y$  sólo puede tomar los valores 0 y 1, si representamos gráficamente la nube de puntos formada por los pares de observaciones de  $Y$  con una de las variables explicativas  $X_i$ , obtendremos puntos situados sobre las rectas  $Y = 1$  e  $Y = 0$ :



Al estimar los parámetros del modelo de probabilidad lineal, estaremos ajustando una recta a la nube de puntos anterior (recta en rojo). El uso de dicha recta para predecir nuevos valores de  $Y$ , i.e., valores de  $P(Y = 1) = Y - u$ , a partir de valores dados de  $X_i$  puede proporcionar valores mayores que 1 o menores que 0 (lo cual está en contradicción con la definición de probabilidad).

4. Finalmente, la expresión  $P(Y = 1) = \beta_1 + \beta_2 X_2 + \dots + \beta_k X_k$  nos dice que la probabilidad del suceso “éxito” viene determinada por una combinación lineal de variables

explicativas. De ello se deduce que  $\frac{\partial P(Y=1)}{\partial X_i} = \beta_i \quad \forall i = 2, 3, \dots, k$ . En otras palabras,

la variación en  $P(Y=1)$  causada por cambios en alguna de las variables explicativas es constante (y, por tanto, independiente del valor actual de dicha variable explicativa), lo cual es una hipótesis muy poco realista. Veamos un ejemplo de ello:

**Ejemplo:** Supongamos que la probabilidad de que una familia adquiera una segunda residencia (suceso “éxito”) viene determinada por su nivel de ingresos netos anuales (IN) según la expresión:

$$P(Y=1) = 0,2 + 0,00001 \cdot IN$$

Usando este modelo, obtendríamos la tabla siguiente:

	Ingresos Netos (euros)	Probabilidad de comprar
Familia 1	10.000	0,3
Familia 2	70.000	0,9

Pues bien, el hecho de que la probabilidad de compra sea una función lineal de la variable IN implicará que un incremento de 10.000 euros en los ingresos netos causará el mismo efecto sobre ambas familias por lo que a la probabilidad de comprar la segunda vivienda se refiere (ambas probabilidades se incrementarán en 0,1). Esto no tiene mucho sentido si tenemos en cuenta que, mientras la primera familia ha visto incrementarse sus ingresos netos en un 100%, para la segunda familia sólo se ha producido un incremento del 14% en su nivel de ingresos netos (por lo que es de esperar que el incremento en la probabilidad de comprar una nueva vivienda sea mayor en la primera de las familias).

## □ Generalización del modelo de probabilidad lineal: modelos Logit y Probit

Como hemos visto, dada una variable dependiente **binaria**  $Y$ , el modelo de regresión lineal  $Y = \beta_1 + \beta_2 X_2 + \dots + \beta_k X_k + u$  presenta serias inconsistencias. Para evitarlas, se han desarrollado modelos no lineales, los cuales tratan de resolver los problemas (3) y (4) anteriores.

La idea consiste en utilizar un modelo de la forma:

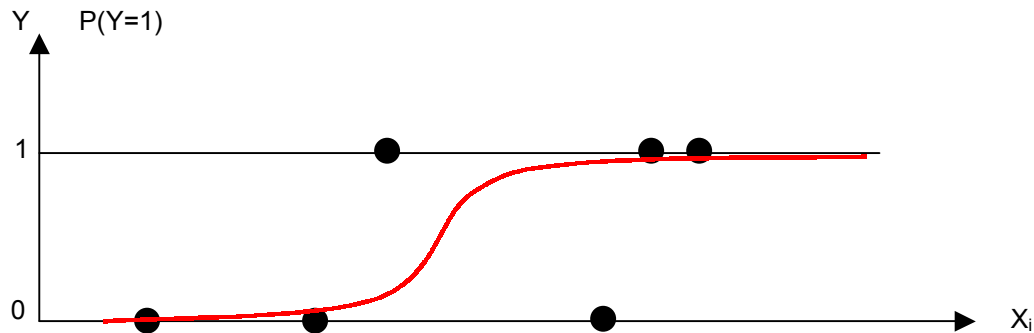
$$Y = f(\beta_1 + \beta_2 X_2 + \dots + \beta_k X_k) + u$$

donde  $f$  es una función real que depende de la expresión lineal  $\beta_1 + \beta_2 X_2 + \dots + \beta_k X_k$  (observar que si  $f$  fuese la función identidad, recuperaríamos el modelo lineal).

Con el nuevo modelo, y razonando de forma similar al caso del modelo lineal, se cumplirá:

$$E[Y] = P(Y=1) = f(\beta_1 + \beta_2 X_2 + \dots + \beta_k X_k)$$

Ahora bien, ¿qué tipo de función  $f$  estamos buscando?: obviamente,  $f$  deberá ser distinta de la función identidad (para evitar el problema (4)). Además, necesitamos una función que esté acotada por los valores 0 y 1 (puesto que su valor coincidirá con el de una probabilidad). En el gráfico siguiente se muestra una idea intuitiva del tipo de función que buscamos para construir nuestro nuevo modelo:



Pues bien, de entre las funciones  $f$  que presentan una forma similar a la de la gráfica, hay dos que son las que se utilizan con mayor frecuencia: la función logística (y que da lugar a los modelos Logit), y la función de distribución de una normal estándar (asociada a los modelos Probit).

### □ El modelo Logit

Como acabamos de ver, una posible solución a las inconsistencias que presentaba el modelo de probabilidad lineal -para explicar el comportamiento de una variable dependiente binaria- es usar un **modelo Logit** de la forma:

$$Y = f(\beta_1 + \beta_2 X_2 + \dots + \beta_k X_k) + u$$

donde  $f$  es la función logística, i.e.:  $f(z) = \frac{\exp(z)}{1 + \exp(z)}$

Por tanto, tendremos que:

$$E[Y] = P(Y = 1) = \frac{\exp(\beta_1 + \beta_2 X_2 + \dots + \beta_k X_k)}{1 + \exp(\beta_1 + \beta_2 X_2 + \dots + \beta_k X_k)}$$

Como hemos comentado al principio de este *math-block*, la estimación en modelos Logit y Probit se realiza mediante el método de Máxima Verosimilitud (MV).

Además, en este tipo de modelos no resulta posible interpretar directamente las estimaciones de los parámetros  $\beta$ , ya que son modelos no lineales. Lo que haremos en la práctica es fijarnos en el signo de los estimadores. Si el estimador es positivo, significará que incrementos en la variable asociada causan incrementos en  $P(Y = 1)$  (aunque desconocemos la magnitud de los mismos). Por el contrario, si el estimador muestra un signo negativo, ello supondrá que incrementos en la variable asociada causarán disminuciones en  $P(Y = 1)$ .

En el modelo Logit se suelen usar otros dos conceptos para profundizar más en la interpretación de los estimadores:

- Se llama **odds** al siguiente cociente de probabilidades:

$$\text{Odds} = \frac{P(Y=1)}{1-P(Y=1)} = \exp(\beta_1 + \beta_2 X_2 + \dots + \beta_k X_k)$$

Tomando logaritmos neperianos en la expresión anterior, obtenemos una expresión lineal para el modelo:

$$\text{Logit}[P(Y=1)] \equiv \ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

Aquí se aprecia claramente que el estimador del parámetro  $\beta_2$  se podrá interpretar como la variación en el **término Logit** (el logaritmo neperiano del cociente de probabilidades) causada por una variación unitaria en la variable  $X_2$  (suponiendo constantes el resto de variables explicativas).

- Cuando se hace referencia al incremento unitario en una de las variables explicativas del modelo, aparece el concepto de **odds-ratio** como el cociente entre los dos **odds** asociados (el obtenido tras realizar el incremento y el anterior al mismo). Así, si suponemos que ha habido un incremento unitario en la variable  $X_i$ , tendremos:

$$\text{Odds-ratio} = \frac{\text{Odds}_2}{\text{Odds}_1} = \exp(\beta_i)$$

De la expresión anterior se deduce que un coeficiente  $\beta_i$  cercano a cero –o, equivalentemente, un **odds-ratio** cercano a uno– significará que cambios en la variable explicativa  $X_i$  asociada no tendrán efecto alguno sobre la variable dependiente  $Y$ .

**Ejemplo:** El archivo **Logit\_Probit.mtw** contiene 32 observaciones para cada una de las variables que se indican a continuación. Cada observación corresponde a un estudiante de la asignatura Econometría:

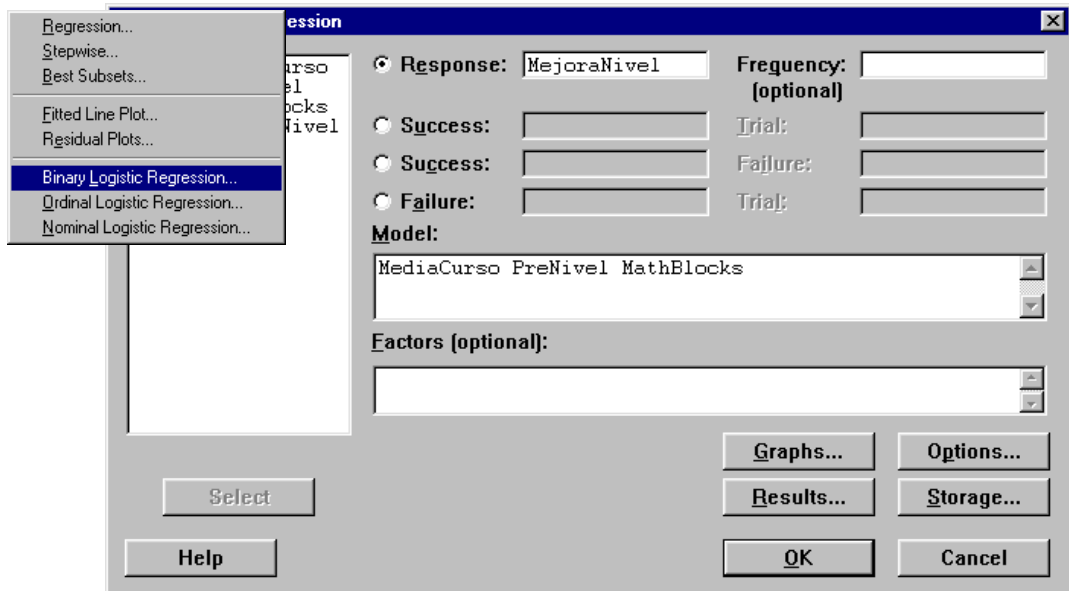
Variable	Descripción
MediaCurso	Contiene la nota media (en una escala de 0 a 10) obtenida por el estudiante en el curso anterior
PreNivel	Contiene un coeficiente (número entero de 10 a 30) que indica el nivel de conocimientos econométricos con que el estudiante comenzó el curso
MathBlocks	Indica si el estudiante ha usado los math-blocks de econometría en su estudio de la asignatura (1 = Sí, 0 = No)
MejoraNivel	Indica si el estudiante ha logrado alcanzar holgadamente los objetivos del curso (1 = Sí, 0 = No)

LOGIT_~1.MTW ***				
	C1	C2	C3	C4
↓	MediaCurso	PreNivel	MathBlocks	MejoraNivel
1	6,32	20	0	0
2	6,78	22	0	0
3	7,56	24	0	0
4	6,84	12	0	0

La idea es explicar, con ayuda de Minitab, el comportamiento de la variable **MejoraNivel** a partir del resto de variables. Para ello usaremos el siguiente modelo Logit:

$$\text{Logit}[P(\text{MejoraNivel} = 1)] = \beta_1 + \beta_2 \cdot \text{MediaCurso} + \beta_3 \cdot \text{PreNivel} + \beta_4 \cdot \text{MathBlocks}$$

Usamos la opción **Stat > Regression > Binary Logistic Regression...** :



Por defecto, Minitab utilizará un modelo Logit. A continuación se muestra el "output":

**Binary Logistic Regression**

Link Function: Logit  
Response Information

Variable	Value	Count
MejoraNi	1	11 (Event)
	0	21
Total		32

**Logistic Regression Table**

Predictor	Coef	StDev	Z	P	Odds Ratio	95% CI	
						Lower	Upper
Constant	-14,434	5,412	-2,67	0,008			
MediaCur	1,4131	0,6315	2,24	0,025	4,11	1,19	14,16
PreNivel	0,0952	0,1416	0,67	0,501	1,10	0,83	1,45
MathBloc	2,379	1,065	2,23	0,025	10,79	1,34	86,94

Log-Likelihood = -12,890  
Test that all slopes are zero: G = 15,404; DF = 3; P-Value = 0,002

**Goodness-of-Fit Tests**

Method	Chi-Square	DF	P
Pearson	27,257	28	0,504
Deviance	25,779	28	0,585
Hosmer-Lemeshow	6,569	8	0,584

**Measures of Association:**  
(Between the Response Variable and Predicted Probabilities)

Pairs	Number	Percent	Summary Measures
Concordant	204	88,3%	Somers' D 0,77
Discordant	26	11,3%	Goodman-Kruskal Gamma 0,77
Ties	1	0,4%	Kendall's Tau-a 0,36
Total	231	100,0%	

La **tabla de regresión logística** muestra los valores estimados para los coeficientes del modelo ( $\beta_1 = -14,434$ ,  $\beta_2 = 1,4131$ ,  $\beta_3 = 0,0952$  y  $\beta_4 = 2,379$ ), junto con sus p-valores asociados (0,008, 0,025, 0,501, y 0,025 respectivamente). Según hemos comentado anteriormente, podemos interpretar los coeficientes  $\beta_2$ ,  $\beta_3$  y  $\beta_4$  como el cambio que se produce en el término Logit al incrementarse en una unidad la variable explicativa asociada. Cuando usamos regresión logística, también nos aparecen los **odds-ratio** (4,11, 1,10 y 10,79 respectivamente).

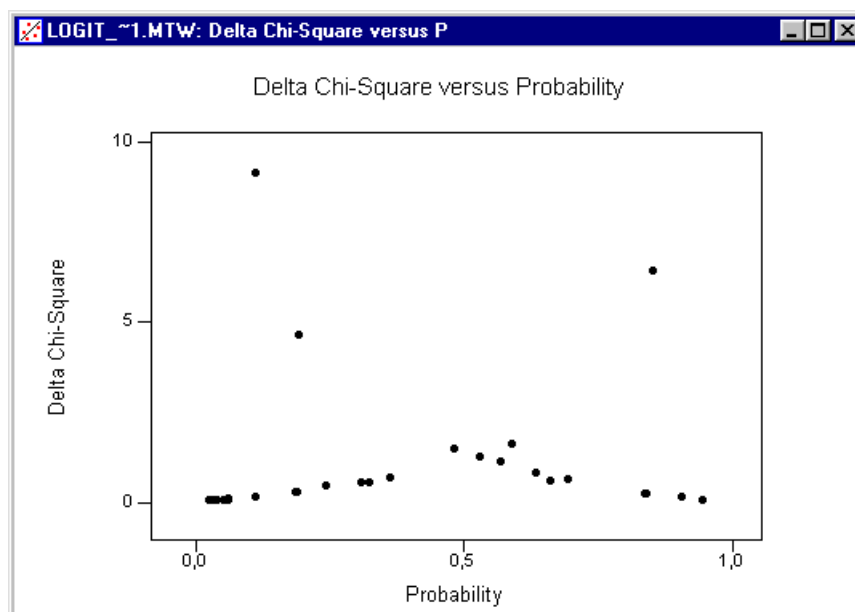
Observar que los p-valores asociados a los coeficientes  $\beta_2$  y  $\beta_4$  son inferiores a 0,05. Por tanto, para un nivel de significación  $\alpha = 0,05$ , rechazaremos la hipótesis nula de que dichos coeficientes son nulos (i.e.: que la variable asociada a los mismos no es relevante en el modelo). Por el contrario, sí parece que la variable **PreNivel** no tendrá un efecto significativo a la hora de explicar el comportamiento de la variable dependiente.

Por lo que se refiere a las variables **MediaCurso** y **MathBlocks**, el hecho de que éstas tengan un coeficiente positivo y un *odds-ratio* bastante mayor a uno, nos hace pensar que cualquier incremento en los niveles de ambas variables (especialmente de la segunda) tendrá un efecto significativo sobre la variable dependiente **MejoraNivel**. Esto significa que aquellos estudiantes que vienen con notas altas del curso anterior, y aquellos estudiantes que hacen uso de los *mathblocks* tienen elevadas probabilidades de alcanzar holgadamente los objetivos del curso.

El **estadístico G** sirve para contrastar la hipótesis nula de que todos los coeficientes asociados con variables explicativas son nulos. Dado que el p-valor obtenido es de 0,002, podemos rechazar dicha hipótesis nula y concluir que, como mínimo, uno de los coeficientes será distinto de cero.

El apartado **Goodness-of-Fit Tests** muestra los p-valores asociados a los contrastes de Pearson, Deviance, y Hosmer-Lemeshow. Dado que dichos p-valores oscilan entre 0,504 y 0,585, no rechazaremos la hipótesis nula de que el modelo se ajusta adecuadamente a las observaciones.

Finalmente, observar en el gráfico siguiente cómo se detecta la existencia de tres observaciones que no son bien explicadas por el modelo:



□ **El modelo Probit**

Otra posible solución a las inconsistencias que presentaba el modelo de probabilidad lineal - para explicar el comportamiento de una variable dependiente binaria- es usar un **modelo Probit** (también llamado **modelo Normit**) de la forma:

$$Y = f(\beta_1 + \beta_2 X_2 + \dots + \beta_k X_k) + u$$

donde f es la función de distribución de una normal estándar, i.e.:

$$f(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp(-t^2 / 2) dt$$

Por tanto, tendremos que:

$$E[Y] = P(Y = 1) = \int_{-\infty}^{\beta_1 + \beta_2 X_2 + \dots + \beta_k X_k} \frac{1}{\sqrt{2\pi}} \exp(-t^2 / 2) dt$$

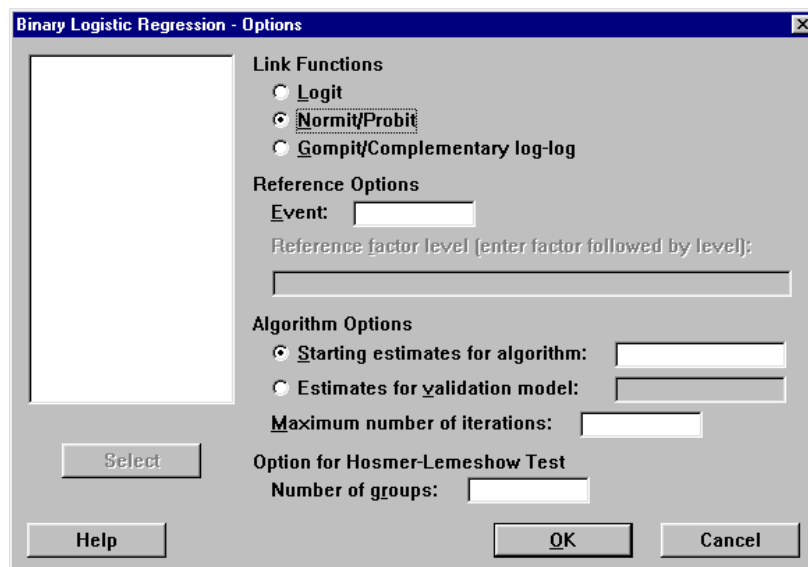
Como hemos comentado al principio de este *math-block*, la estimación en modelos Logit y Probit se realiza mediante el método de Máxima Verosimilitud (MV).

Además, en este tipo de modelos no resulta posible interpretar directamente las estimaciones de los parámetros  $\beta$ , ya que son modelos no lineales. Lo que haremos en la práctica es fijarnos en el signo de los estimadores. Si el estimador es positivo, significará que incrementos en la variable asociada causan incrementos en  $P(Y = 1)$  (aunque desconocemos la magnitud de los mismos). Por el contrario, si el estimador muestra un signo negativo, ello supondrá que incrementos en la variable asociada causarán disminuciones en  $P(Y = 1)$ .

**Ejemplo:** Volviendo al ejemplo anterior, utilizaremos ahora el siguiente modelo Probit para explicar el comportamiento de la variable dependiente binaria **MejoraNivel**:

$$Pr obit[P(MejoraNivel = 1)] = \beta_1 + \beta_2 \cdot MediaCurso + \beta_3 \cdot Pr eNivel + \beta_4 \cdot MathBlocks$$

Los pasos a realizar con Minitab son análogos al ejemplo anterior. Eso sí, deberemos especificar en este caso que el modelo a usar es el Normit (o Probit). Usaremos para ello el menú **options**:



Si nos fijamos en el “output” siguiente, veremos que los coeficientes obtenidos para el modelo Probit son consistentes con los que habíamos obtenido para el modelo Logit (en el sentido de que ambos tienen el mismo signo y, por tanto, la interpretación de unos es coherente con la de los otros):

**Binary Logistic Regression**

Link Function: Normit  
Response Information

Variable	Value	Count	
MejoraNi	1	11	(Event)
	0	21	
Total		32	

**Logistic Regression Table**

Predictor	Coef	StDev	Z	P
Constant	-8,265	2,821	-2,93	0,003
MediaCur	0,8129	0,3449	2,36	0,018
PreNivel	0,05173	0,08119	0,64	0,524
MathBloc	1,4263	0,5870	2,43	0,015

Log-Likelihood = -12,819  
Test that all slopes are zero: G = 15,546; DF = 3; P-Value = 0,001

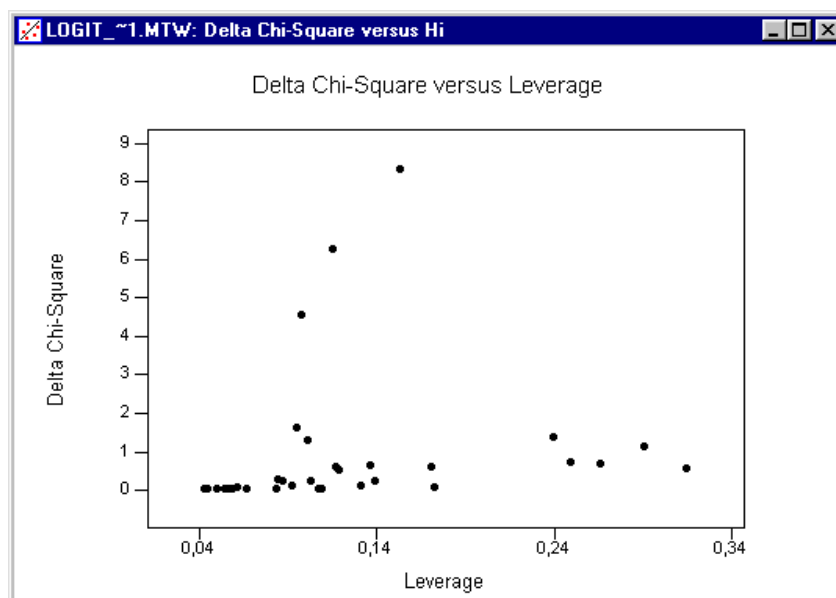
**Goodness-of-Fit Tests**

Method	Chi-Square	DF	P
Pearson	26,252	28	0,559
Deviance	25,638	28	0,593
Hosmer-Lemeshow	6,909	8	0,547

Measures of Association:  
(Between the Response Variable and Predicted Probabilities)

Pairs	Number	Percent	Summary Measures	
Concordant	204	88,3%	Somers' D	0,77
Discordant	26	11,3%	Goodman-Kruskal Gamma	0,77
Ties	1	0,4%	Kendall's Tau-a	0,36
Total	231	100,0%		

En este caso, el resultado del “output” proveniente del modelo Probit es análogo al que habíamos realizado para el modelo Logit. Además, también detectamos las tres observaciones que el modelo no explica convenientemente:



**CASOS PRÁCTICOS CON SOFTWARE**

□ **Ejemplo de regresión binaria mediante un modelo Logit**

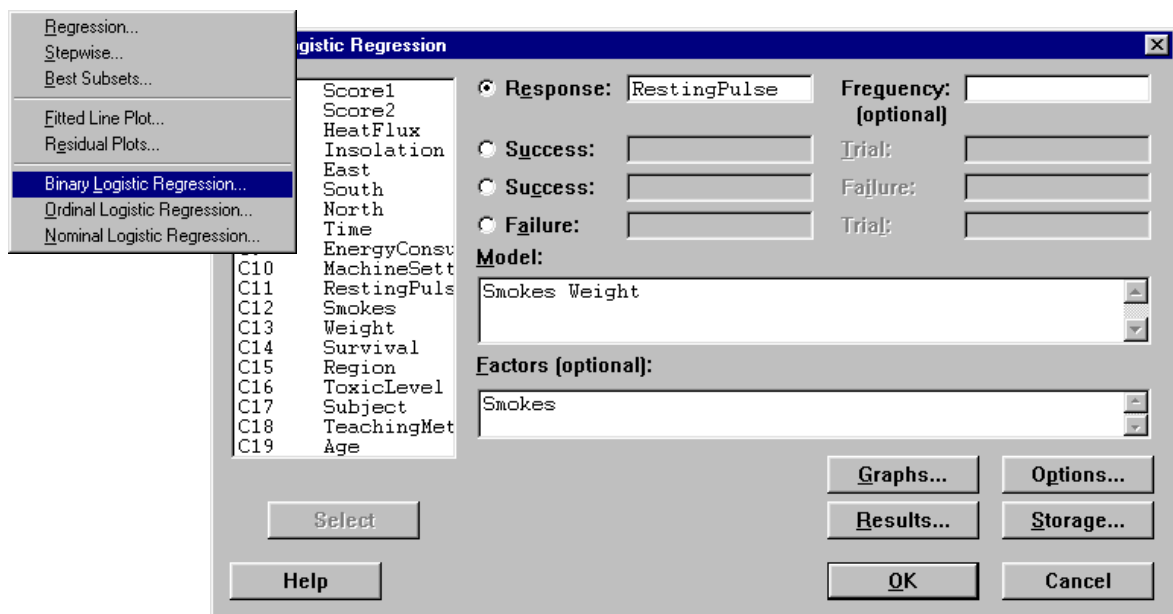
Pretendemos investigar el efecto del tabaco (si se fuma o no) y del peso sobre la frecuencia de pulsaciones que una persona tiene en reposo. Hemos categorizado la variable dependiente, **RestingPulse**, en dos niveles: nivel bajo (*Low*) y nivel alto (*High*). Los datos (observaciones) se encuentran en el archivo **Exh\_regr.mtw**:

	C10	C11-T	C12-T	C13	C14	C15	C16	C17-T	C18-T
↓	MachineSetting	RestingPulse	Smokes	Weight	Survival	Region	ToxicLevel	Subject	TeachingMethod
1	11,15	Low	No	140	1	1	62,00	math	discuss
2	15,70	Low	No	145	1	2	46,00	science	discuss
3	18,90	Low	Yes	160	2	1	48,50	science	discuss
4	19,40	Low	Yes	190	3	2	32,00	math	lecture
5	21,40	Low	No	155	2	1	63,50	math	discuss
6	21,70	Low	No	165	1	1	41,25	science	lecture
7	25,00	High	Yes	150	2	2	40,00	math	lecture

Emplearemos para ello el siguiente modelo Logit:

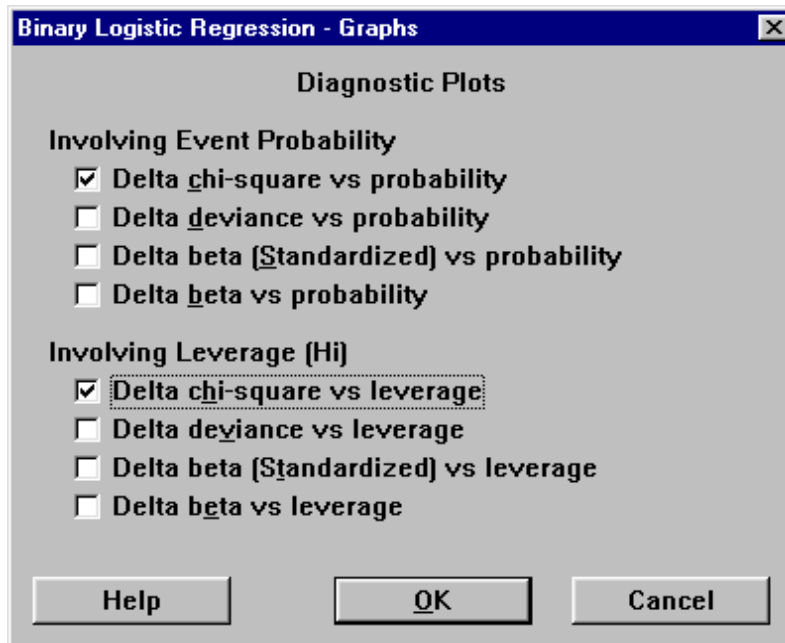
$$\text{Logit}[P(\text{RestingPulse} = \text{alto})] = \beta_1 + \beta_2 \cdot \text{Smokes} + \beta_3 \cdot \text{Weight}$$

Usamos la opción **Stat > Regression > Binary Logistic Regression...** :



Observar que el modelo debe incluir las dos variables explicativas (**Smokes** y **Weight**). Así mismo, hemos especificado que la variable **Smokes** es un factor (Minitab denomina **factors** a las variables explicativas categóricas, y **covariates** a las variables explicativas cuantitativas).

Pediremos también que el “output” incluya un par de gráficos:



A continuación se muestra el “output” generado por Minitab:

**Results for: Exh\_regr.MTW**

**Binary Logistic Regression: RestingPulse versus Smokes; Weight**

Link Function: Logit

Response Information

Variable	Value	Count
RestingP	Low	70 (Event)
	High	22
	Total	92

**Logistic Regression Table**

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI	
						Lower	Upper
Constant	-1,987	1,679	-1,18	0,237			
Smokes							
Yes	-1,1930	0,5530	-2,16	0,031	0,30	0,10	0,90
Weight	0,02502	0,01226	2,04	0,041	1,03	1,00	1,05

Log-Likelihood = -46,820  
 Test that all slopes are zero: G = 7,574; DF = 2; P-Value = 0,023

**Goodness-of-Fit Tests**

Method	Chi-Square	DF	P
Pearson	40,848	47	0,724
Deviance	51,201	47	0,312
Hosmer-Lemeshow	4,745	8	0,784

La **tabla de regresión logística** muestra los valores estimados para los coeficientes del modelo ( $\beta_1 = -1,987$ ,  $\beta_2 = -1,1930$ , y  $\beta_3 = 0,02502$ ), junto con sus p-valores asociados (0,237, 0,031, y 0,041 respectivamente). Según hemos comentado anteriormente, podemos interpretar los coeficientes  $\beta_2$  y  $\beta_3$  como el cambio que se produce en el término Logit al incrementarse en una unidad la variable explicativa asociada. Cuando usamos regresión logística, también nos aparecen los **odds-ratio** (0,30 y 1,03 respectivamente).

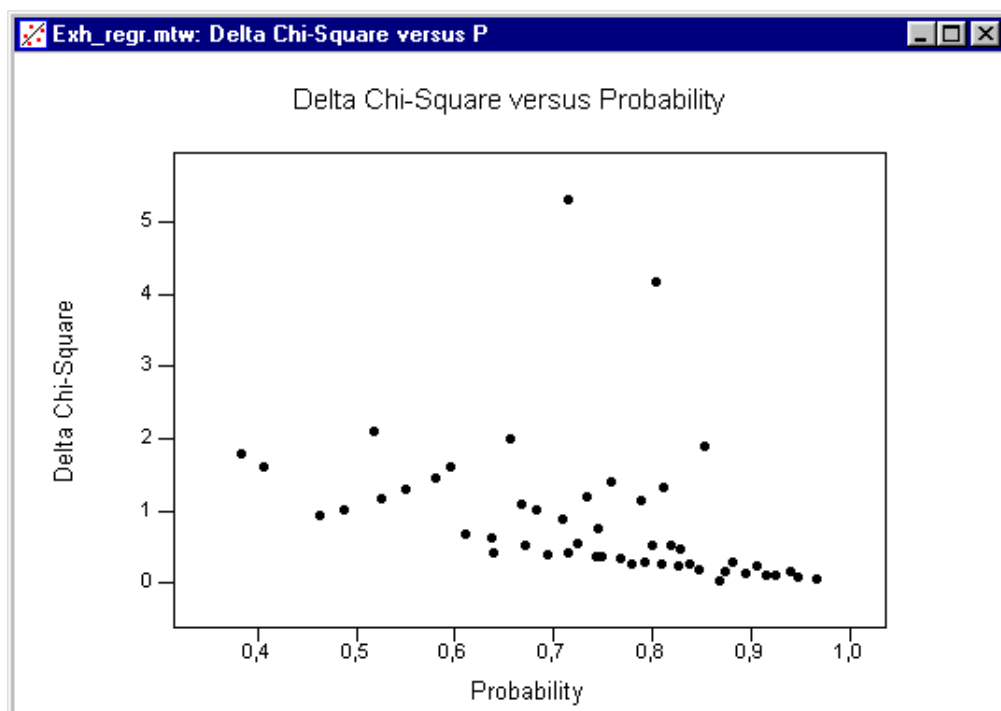
Observar que los p-valores asociados a los coeficientes  $\beta_2$  y  $\beta_3$  son inferiores a 0,05. Por tanto, para un nivel de significación  $\alpha = 0,05$ , rechazaremos la hipótesis nula de que dichos coeficientes son nulos (i.e.: que la variable asociada a los mismos no es relevante en el modelo).

Si bien no hay indicios de que alguno de los coeficientes  $\beta_2$  o  $\beta_3$  sea nulo, el hecho de que el **odds-ratio** asociado a la variable **Weight** valga aproximadamente uno, induce a pensar que un aumento unitario en el peso de un individuo no causará un efecto apreciable sobre su frecuencia de pulsaciones en reposo. Por lo que se refiere a la variable **Smokes**, el hecho de que ésta tenga un coeficiente negativo y un **odds-ratio** de 0,30 (diferente de uno), nos hace pensar que aquellos individuos que fuman tienden a tener una mayor frecuencia de pulsaciones en reposo que los que no fuman.

El **estadístico G** sirve para contrastar la hipótesis nula de que todos los coeficientes asociados con variables explicativas son nulos. Dado que el p-valor obtenido es de 0,023, podemos rechazar dicha hipótesis nula y concluir que, como mínimo, uno de los coeficientes será distinto de cero.

El apartado **Goodness-of-Fit Tests** muestra los p-valores asociados a los contrastes de Pearson, Deviance, y Hosmer-Lemeshow. Dado que dichos p-valores oscilan entre 0,312 y 0,724, no rechazaremos la hipótesis nula de que el modelo se ajusta adecuadamente a las observaciones.

Finalmente, observar en el gráfico siguiente cómo se detecta la existencia de dos observaciones que no son bien explicadas por el modelo:



□ **Ejemplo de regresión binaria mediante un modelo Probit**

En el ejemplo anterior, hubiésemos podido emplear el siguiente modelo Probit:

$$Pr\text{obit}[P(\text{RestingPulse} = \text{alto})] = \beta_1 + \beta_2 \cdot \text{Smokes} + \beta_3 \cdot \text{Weight}$$

En tal caso, el “output” hubiese sido el siguiente:

Binary Logistic Regression				
Link Function: Normit				
Response Information				
Variable	Value	Count		
RestingP	Low	70	(Event)	
	High	22		
	Total	92		
Logistic Regression Table				
Predictor	Coef	StDev	Z	P
Constant	-1,2011	0,9764	-1,23	0,219
Smokes				
Yes	-0,7038	0,3250	-2,17	0,030
Weight	0,015085	0,007025	2,15	0,032
Log-Likelihood = -46,734				
Test that all slopes are zero: G = 7,746; DF = 2; P-Value = 0,021				
Goodness-of-Fit Tests				
Method	Chi-Square	DF	P	
Pearson	40,598	47	0,733	
Deviance	51,029	47	0,318	
Hosmer-Lemeshow	5,845	8	0,665	
Measures of Association: (Between the Response Variable and Predicted Probabilities)				
Pairs	Number	Percent	Summary Measures	
Concordant	1046	67,9%	Somers' D	0,38
Discordant	462	30,0%	Goodman-Kruskal Gamma	0,39
Ties	32	2,1%	Kendall's Tau-a	0,14
Total	1540	100,0%		

Nuevamente, podemos apreciar que los coeficientes obtenidos en ambos modelos (Logit y Probit) son consistentes (tienen el mismo signo en ambos modelos y, por tanto, su interpretación es análoga).

Asimismo, el resto del “output” del modelo Probit es coherente con el análisis que habíamos realizado para el modelo Logit:

La **tabla de regresión logística** muestra los valores estimados para los coeficientes del modelo ( $\beta_1 = -1,2011$ ,  $\beta_2 = -0,7038$ , y  $\beta_3 = 0,015085$ ), junto con sus p-valores asociados (0,219, 0,030, y 0,032 respectivamente). Según hemos comentado anteriormente, podemos interpretar los coeficientes  $\beta_2$  y  $\beta_3$  como el cambio que se produce en el término Probit al incrementarse en una unidad la variable explicativa asociada.

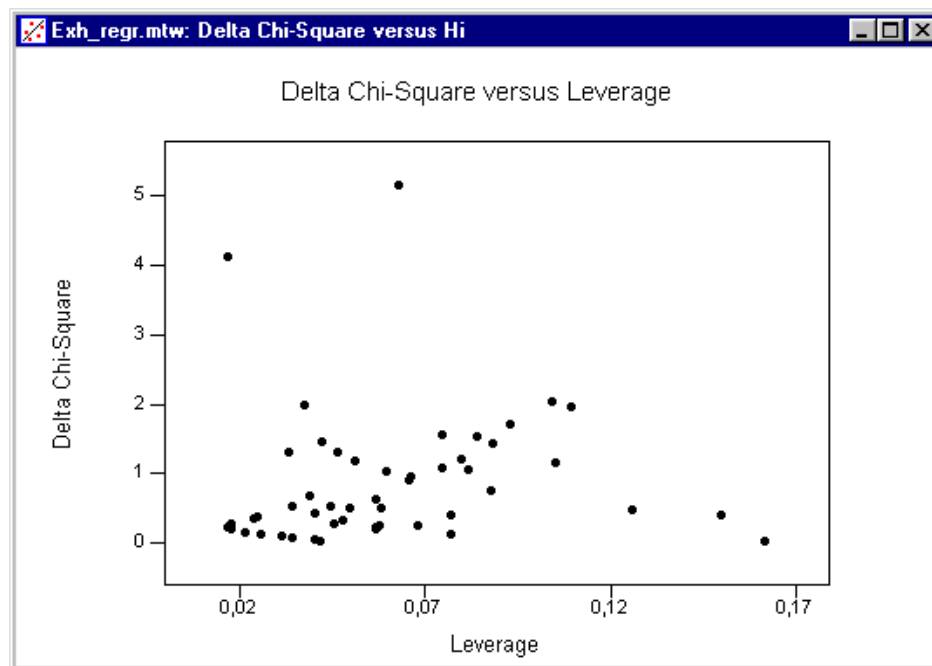
Observar que los p-valores asociados a los coeficientes  $\beta_2$  y  $\beta_3$  son inferiores a 0,05. Por tanto, para un nivel de significación  $\alpha = 0,05$ , rechazaremos la hipótesis nula de que dichos coeficientes son nulos (i.e.: que la variable asociada a los mismos no es relevante en el modelo).

Si bien no hay indicios de que alguno de los coeficientes  $\beta_2$  o  $\beta_3$  sea nulo, el hecho de que el coeficiente asociado a la variable **Weight** valga aproximadamente cero, induce a pensar que un aumento unitario en el peso de un individuo no causará un efecto apreciable sobre su frecuencia de pulsaciones en reposo. Por lo que se refiere a la variable **Smokes**, el hecho de que ésta tenga un coeficiente negativo mayor que uno, nos hace pensar que aquellos individuos que fuman tienden a tener una mayor frecuencia de pulsaciones en reposo que los que no fuman.

El **estadístico G** sirve para contrastar la hipótesis nula de que todos los coeficientes asociados con variables explicativas son nulos. Dado que el p-valor obtenido es de 0,021, podemos rechazar dicha hipótesis nula y concluir que, como mínimo, uno de los coeficientes será distinto de cero.

El apartado **Goodness-of-Fit Tests** muestra los p-valores asociados a los contrastes de Pearson, Deviance, y Hosmer-Lemeshow. Dado que dichos p-valores oscilan entre 0,318 y 0,733, no rechazaremos la hipótesis nula de que el modelo se ajusta adecuadamente a las observaciones.

Finalmente, en el gráfico siguiente se aprecian claramente las dos observaciones que no son bien explicadas por el modelo:



## BIBLIOGRAFÍA

---

- [1] Artís, M.; Suriñach, J.; et al (2002): “Econometría”. Ed. Fundació per a la Universitat Oberta de Catalunya. Barcelona.
- [2] Doran, H. (1989): “Applied Regression Analysis in Econometrics”. Ed. Marcel Dekker, Inc. ISBN: 0-8247-8049-3
- [3] Frone, M.R. (1997): “Regression Models for Discrete and Limited Dependent Variables”. [http://www.aom.pace.edu/rmd/1997\\_forum\\_regression\\_models.html](http://www.aom.pace.edu/rmd/1997_forum_regression_models.html)
- [4] Gujarati, D. (1997): “Econometría básica”. McGraw-Hill. ISBN 958-600-585-2
- [5] Johnston, J. (2001): “Métodos de econometría”. Ed. Vicens Vives. Barcelona. ISBN 84-316-6116-X
- [6] Kennedy, P. (1998): “A Guide to Econometrics”. Ed. MIT Press. ISBN: 0262611406
- [7] Novalés, A. (1993): “Econometría”. McGraw-Hill. ISBN 84-481-0128-6
- [8] Pulido, A. (2001): “Modelos econométricos”. Ed. Pirámide. Madrid. ISBN 84-368-1534-3
- [9] Uriel, E. (1990): “Econometría: el modelo lineal”. Ed. AC. Madrid. ISBN 84-7288-150-4
- [10] Wooldridge, J. (2001): “Introducción a la Econometría: un enfoque moderno”. Ed. Thomson Learning. ISBN: 970-686-054-1

## ENLACES

---

- ❑ <http://www2.chass.ncsu.edu/garson/pa765/logit.htm>  
Log-Linear, Logit, and Probit Models
- ❑ <http://www.la.utexas.edu/research/faculty/dpowers/book/htmlbook/contents.html>  
Libro on-line: Statistical Methods for Categorical Data Analysis
- ❑ <http://www.feweb.vu.nl/econometriclinks/index.html>  
The Econometrics Journal On-Line
- ❑ <http://www.elsevier.com/hes/books/02/menu02.htm>  
Libro on-line: Handbook of Econometrics Vols. 1-5
- ❑ <http://elsa.berkeley.edu/users/mcfadden/discrete.html>  
Libro on-line: Structural Analysis of Discrete Data and Econometric Applications
- ❑ [http://www.oswego.edu/~kane/econometrics/stud\\_resources.htm](http://www.oswego.edu/~kane/econometrics/stud_resources.htm)  
Online Resources for Econometric Students
- ❑ <http://www.econ.uiuc.edu/~morillo/links.html>  
Econometric Sources: a collection of links in econometrics and computing. University of Illinois