



ugr

Universidad  
de Granada

---

# Métodos de Bondad de Ajuste en Regresión Logística

MÁSTER OFICIAL EN ESTADÍSTICA APLICADA

TRABAJO FIN DE MÁSTER

---

*Autor:*

Tania Iglesias Cabo

*Dirigido:*

Manuel Escabias Macucha

Ana María Aguilera del Pino

Curso académico 2012/2013



# Índice general

<b>1. Introducción</b>	<b>3</b>
<b>2. El modelo de regresión logística</b>	<b>9</b>
2.1. Formulación e interpretación . . . . .	10
2.1.1. Formulación . . . . .	10
2.1.2. Interpretación . . . . .	11
2.2. Estimación . . . . .	12
2.2.1. Estimación del modelo simple binario . . . . .	13
2.2.2. Estimación del modelo simple binomial . . . . .	15
<b>3. Bondad de ajuste en regresión logística</b>	<b>19</b>
3.1. Test basados en patrones de las covariables . . . . .	20
3.1.1. Estadístico basado en la Devianza $D$ . . . . .	20
3.1.2. Estadístico Chi Cuadrado de Pearson $\chi^2$ . . . . .	23
3.2. Test basados en probabilidades estimadas . . . . .	24
3.2.1. Estadístico de Hosmer-Lemeshow $C_g$ . . . . .	25
3.2.2. Estadístico de Hosmer-Lemeshow $H_g$ . . . . .	26
3.3. Test Score . . . . .	28
3.3.1. Estadístico Score de Brown $\hat{B}$ . . . . .	28
3.3.2. Estadístico Score de Stukel $\hat{S}_{ST}$ . . . . .	31

3.3.3. Estadístico de Tsiatis $T$ . . . . .	32
3.4. Test basados en residuos suavizados . . . . .	33
3.4.1. Estadístico de le Cessie y Van Houwelingen $\hat{T}_{lc}$ . . . . .	34
3.5. Medidas tipo $R^2$ . . . . .	36
<b>4. Software</b>	<b>39</b>
4.1. R . . . . .	39
4.2. SPSS . . . . .	46
4.3. Stata . . . . .	47
<b>5. Aplicación con datos reales</b>	<b>51</b>
5.1. Descripción de los datos . . . . .	51
5.2. Modelo de regresión logística simple . . . . .	54
5.2.1. Ajuste y bondad del modelo con R . . . . .	54
5.2.2. Ajuste y bondad del modelo con SPSS . . . . .	58
5.2.3. Ajuste y bondad del modelo con Stata . . . . .	62
5.3. Modelo de regresión logística múltiple . . . . .	66
5.3.1. Ajuste y bondad del modelo con R . . . . .	66
5.3.2. Ajuste y bondad del modelo con SPSS . . . . .	69
5.3.3. Ajuste y bondad del modelo con Stata . . . . .	70
5.4. Conclusiones . . . . .	73

# Capítulo 1

## Introducción

Con este trabajo se pretende hacer una descripción de distintos test estadísticos utilizados para medir la bondad de ajuste del modelo de regresión logística. El modelo de regresión logística se enmarca dentro de los modelos de respuesta discreta, válidos para modelar la relación entre una respuesta discreta y un conjunto de variables independientes, o explicativas. Habitualmente, en marcos como el epidemiológico y el biosanitario, se trabaja con variables respuesta que suelen representar la presencia o ausencia de enfermedad y variables explicativas que identifican posibles factores de riesgo. La técnica más ampliamente utilizada para modelar respuestas discretas es la regresión logística, en la que las variables explicativas pueden ser tanto de tipo cuantitativo, como de tipo cualitativo. La popularidad de este tipo de técnica radica principalmente en su disponibilidad en los distintos software, así como en su facilidad en cuanto a la interpretación de la exponencial de los parámetros del modelo, en términos de los cocientes de ventajas u odds ratios (OR).

Una vez se ha ajustado un modelo, tiene sentido plantearse cómo de correcto es dicho ajuste a los datos. Se considera que un modelo no presenta un

buen ajuste si la variabilidad residual es grande, sistemática o no se corresponde con la variabilidad descrita por el modelo (Ryan [1997]). Un modelo de regresión logística puede resultar inadecuado por diferentes razones. Una de ellas es la invalidez de la componente lineal del modelo, esta situación es frecuente cuando ciertas variables predictoras o términos de interacción no son incluidos en el modelo debiendo ser incluidos, o también cuando no se realizan transformaciones en ellas que permiten mejorar el ajuste a los datos. La existencia de observaciones influyentes o outliers puede ser también determinante en un mal ajuste. Otra de las situaciones habituales en un mal ajuste se tiene cuando la función de enlace no es la apropiada. Como la transformación logística es la más utilizada en la práctica, suele elegirse obviando que en ocasiones un modelo probit o de valores extremos podría proporcionar mejores resultados. Un escenario de mal ajuste es posible también cuando no se cumple la hipótesis de distribución binomial de la variable respuesta, que puede ser debido a que las proporciones observadas no sean independientes, o a la existencia de sobredispersión o infradispersión (Collett [1991]) provocando una estimación de la varianza mayor o inferior que la que debería ser.

Un modelo se dice que presenta un buen ajuste a los datos si los valores predichos por él reflejan de forma adecuada a los valores observados. Si el modelo presenta un mal ajuste, éste no puede ser utilizado para extraer conclusiones ni efectuar predicciones. Una forma de medir la adecuación de un modelo es proporcionando medidas globales de bondad de ajuste a través de test estadísticos contruidos a tal fin. Lo anterior es válido para cualquier modelo de regresión, pero en el caso particular de la regresión logística no existe uniformidad en cuanto al test a utilizar.

Así, los objetivos de este trabajo son:

1. Ampliar los conocimientos adquiridos durante el curso teórico realizado, identificando test estadísticos, fuera de los más usuales y populares, que permitan medir la bondad de ajuste de un modelo de regresión logística, determinando las ventajas y limitaciones de cada uno, tanto desde un punto de vista teórico como práctico. Se ha trabajado tomando como base los métodos descritos en la tesis de Hallet (Hallet [1999]).
2. Revisar la disponibilidad de los métodos de bondad de ajuste en distintos paquetes estadísticos.
3. Aplicar los distintos métodos a un conjunto de datos reales, poniendo de manifiesto las diferencias entre los diferentes test y las diferencias entre los distintos programas.

En cuanto a la organización de este trabajo, se ha optado por una división en los siguientes capítulos:

- Capítulo 1: Introducción. En este capítulo se proporciona una visión general y resumida del trabajo, planteando la temática sobre el que versará, indicando los objetivos y organización del mismo.
- Capítulo 2: El modelo de regresión logística. Se presenta desde un punto de vista teórico la formulación, interpretación y estimación de los parámetros de un modelo de regresión logística.
- Capítulo 3: Bondad de ajuste en regresión logística. Se introducen desde un punto de vista teórico diversos métodos estadísticos que permiten el estudio de la bondad de ajuste de un modelo de regresión logística.
- Capítulo 4: Software. Se detallan y comparan los métodos de bondad de ajuste implementados por tres programas destinados al análisis de datos.

- Capítulo 5: Aplicación a datos reales. En este capítulo se aplican los distintos métodos a un conjunto de datos reales, poniendo de relieve las diferencias entre ellos y entre los distintos software.







## Capítulo 2

# El modelo de regresión logística

Los modelos de regresión logística son una herramienta que permite explicar el comportamiento de una variable respuesta discreta (binaria o con más de dos categorías) a través de una o varias variables independientes explicativas de naturaleza cuantitativa y/o cualitativa. Según el tipo de variable respuesta estaremos hablando de regresión logística binaria (variable dependiente con 2 categorías), o de regresión logística multinomial (variable dependiente con más de 2 categorías), pudiendo ser esta última de respuesta nominal u ordinal. Los modelos de respuesta discreta son un caso particular de los modelos lineales generalizados formulados por Nelder y Wedderburn en 1972 (Nelder and Wedderburn [1972]), al igual que los modelos de regresión lineal o el análisis de la varianza. Para un estudio minucioso de este tipo de modelos puede consultarse el libro de McCullagh y Nelder (McCullagh and Nelder [1989]).

A continuación plantearemos la formulación genérica de un modelo de regresión logística binaria, así como la interpretación y estimación de los parámetros, aunque para un estudio exhaustivo son recomendables la consulta de otros materiales indicados en la bibliografía (Agresti [2002]; Hosmer

and Lemeshow [1989];Thompson [2007];Power and Xie [2000] ;Kleinbaum [1994];Selvin [1996];Aycaguer and Utra;Venables and Ripley [2003] ).

## 2.1. Formulación e interpretación

### 2.1.1. Formulación

Supongamos que tenemos una variable respuesta o dependiente  $Y$  que toma dos valores, que habitualmente serán  $Y = 1$  (suele indicar presencia de cierta característica u ocurrencia de cierto suceso) e  $Y = 0$  (ausencia de característica o no observación del suceso). Denotemos por  $R$  el número de variables independientes del modelo representadas por  $X = (X_1, X_2, \dots, X_R)'$ . La formulación genérica del modelo de regresión logística para modelar la probabilidad de ocurrencia de un suceso sería  $Y = p_x + \varepsilon$  donde  $\varepsilon$  es el término de error,  $p_x$  es la probabilidad de que la respuesta  $Y$  tome el valor 1 para el valor observado  $x$  y se modeliza como:

$$P(Y = 1|X = x) = p_x = \frac{\exp\left(\beta_0 + \sum_{r=1}^R \beta_r x_r\right)}{1 + \exp\left(\beta_0 + \sum_{r=1}^R \beta_r x_r\right)}$$

siendo  $x = (x_1, x_2, \dots, x_R)'$  un valor observado de las variables explicativas. Por tanto,  $1 - p_x$  indicará la probabilidad de que  $Y$  tome el valor 0.

Si aplicamos una transformación logit a la ecuación anterior, obtenemos un modelo de regresión lineal que facilitará la posterior interpretación del modelo:

$$\text{logit}(p_x) = \log\left(\frac{p_x}{1 - p_x}\right) = \beta_0 + \sum_{r=1}^R \beta_r x_r$$

El objetivo de este trabajo es el estudio de diferentes métodos de análisis de la bondad del ajuste en regresión logística. Tales métodos se basan en comparar las observaciones de la respuesta con las predicciones hechas por el modelo, independientemente de las variables explicativas utilizadas para tales predicciones. Por ello, y con objeto de hacer más fácil la lectura del trabajo, centraremos el desarrollo del mismo, sin pérdida de generalidad, y a efectos del estudio de la bondad del ajuste, en el modelo de regresión logística simple, que es aquel que tiene una única variable explicativa.

Así, considerando  $Y$  la variable respuesta,  $X$  la variable predictora y  $x$  un valor observado de ésta, se formula el modelo logístico simple como  $Y = p_x + \varepsilon$  donde ahora:

$$P(Y = 1|X = x) = p_x = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}.$$

Utilizando la función logit obtendríamos el modelo lineal:

$$\text{logit}(p_x) = \log\left(\frac{p_x}{1 - p_x}\right) = \beta_0 + \beta_1 x$$

### 2.1.2. Interpretación

En la formulación del modelo de regresión logística simple, existen dos coeficientes  $\beta_0$  y  $\beta_1$  que interpretaremos a continuación en términos de cocientes de ventajas. El cociente de ventajas de respuesta  $Y = 1$  para  $x_1$  y  $x_2$  dos valores distintos de  $X$  se define como:

$$\theta(x_1, x_2) = \frac{\frac{p_{x_1}}{1 - p_{x_1}}}{\frac{p_{x_2}}{1 - p_{x_2}}}.$$

A partir de la formulación del modelo se tiene de manera inmediata (ver por ejemplo Ryan [1997]) que:

- $\beta_0$ , término constante del modelo o intercept: se corresponde con el logaritmo de la ventaja de respuesta  $Y = 1$  frente a la respuesta  $Y = 0$

para una observación con valor nulo en la variable explicativa, es decir, cuando la respuesta es independiente de las variable explicativa. Por tanto, la exponencial de la constante  $e^{\beta_0}$  será la ventaja de respuesta  $Y = 1$  para un individuo con  $X = 0$ .

- $\beta_1$  o parámetro *slope*: cuya exponencial es el cociente de ventaja de respuesta  $Y = 1$  u odds ratio dado por dos observaciones de la variable explicativa que se diferencian en una unidad. Si la variable predictora es de tipo continuo, la exponencial del parámetro  $\beta_1$  es el cociente de ventajas para un incremento de una unidad en la variable explicativa. Así, la ventaja de respuesta  $Y = 1$  queda multiplicada por dicha exponencial al aumentar en una unidad la variable explicativa.

## 2.2. Estimación

Existen dos formas de estimación del modelo logístico según sean las observaciones disponibles, dicho de otro modo, según el patron de las covariables (covariate pattern) que no es más que cada combinación de valores de las variables explicativas en el modelo múltiple, o las observaciones de la variable explicativa en el modelo simple. Hay que distinguir dos situaciones diferentes. Supongamos que disponemos de una muestra de tamaño  $N$  de la variable respuesta  $Y$ , puede ocurrir:

1. Que exista en cada valor de la variable/s explicativa/s varias observaciones de la respuesta  $y_j$ . Si denotamos por  $J$  el número de valores distintos de la variable, entonces para cada  $x_j$  ( $j = 1, \dots, J$ ),  $n_j$  es el número de observaciones de la respuesta, en cuyo caso llamaremos  $y_j$  al número de éxitos en las  $n_j$  observaciones de la respuesta en cada valor de la variable explicativa. En este caso se habla de da-

tos agrupados y  $J < N = \sum_{j=1}^J n_j$ , por tanto, el número de patrones sería inferior al número de observaciones. En esta circunstancia la respuesta se considera agrupada resultando una distribución binomial  $y_j \rightsquigarrow B(n_j, p_j)$ ,  $j = 1, \dots, J$  por lo que este análisis recibe el nombre de regresión logística binomial.

2. Que exista en cada valor de la variable/s explicativa/s  $x_j$  una única observación de la respuesta  $y_j$ , es decir, los datos se presentan no agrupados  $(x_1, y_1), \dots, (x_N, y_N)$ . Si denotamos por  $J$  el número de valores distintos de la variable  $X$ , se tiene que  $J = N$ , es decir, el número de patrones es igual al número de observaciones. Esta situación suele darse al trabajar con variables explicativas continuas, resultando en una regresión logística binaria, puesto que la respuesta se considera agrupada resultando una distribución bernouilli  $y_j \rightsquigarrow B(1, p_j)$ .

### 2.2.1. Estimación del modelo simple binario

En un modelo de regresión logística simple, es decir, con una única variable explicativa, los dos parámetros desconocidos  $\beta_0$  y  $\beta_1$  son estimados usando el método de máxima verosimilitud, que consiste en proporcionar la estimación que otorgue máxima probabilidad o verosimilitud a los datos observados.

En el escenario de regresión logística binaria simple descrita anteriormente, y suponiendo las observaciones independientes, la función de verosimilitud es de la forma:

$$L(\beta_0, \beta_1) = \prod_{j=1}^N p_j^{y_j} (1 - p_j)^{1-y_j}$$

para  $y_1, \dots, y_N$  observaciones de  $Y$  ( $y_j \in \{0, 1\}$ ),  $p_j = \frac{\exp(\beta_0 + \beta_1 x_j)}{1 + \exp(\beta_0 + \beta_1 x_j)}$ ,  $j = 1, \dots, N$  y  $x_1, \dots, x_N$  observaciones de  $X$ .

La estimación de los dos coeficientes requiere maximizar la función de verosimilitud, o equivalentemente, maximizar su logaritmo:

$$\log(L(\beta_0, \beta_1)) = \sum_{j=1}^N (y_j \log(p_j) + (1 - y_j) \log(1 - p_j))$$

Derivando respecto a cada uno de los  $(\beta_0, \beta_1)$  e igualando a cero obtenemos las ecuaciones de verosimilitud:

$$\sum_{j=1}^N (y_j - p_j) = 0$$

y

$$\sum_{j=1}^N x_j (y_j - p_j) = 0$$

Los estimadores MV de un modelo logit siempre existen y son únicos (salvo en ciertos casos de separación completa) debido a la concavidad de la log-verosimilitud. Es necesario entonces para la existencia de estos estimadores que exista cierto solapamiento en los datos (Santner and Duffy [1986]). Las ecuaciones obtenidas no son lineales en los parámetros, de aquí que requieran métodos iterativos como el de Newton-Raphson para su resolución. La fórmula iterativa de resolución de las ecuaciones de verosimilitud es (ver Ryan [1997]):

$$\begin{pmatrix} \beta_0^{(t)} \\ \beta_1^{(t)} \end{pmatrix} = \begin{pmatrix} \beta_0^{(t-1)} \\ \beta_1^{(t-1)} \end{pmatrix} + \begin{pmatrix} \sum_{j=1}^N p_j^{(t-1)}(1 - p_j^{(t-1)}) & \sum_{j=1}^N x_j p_j^{(t-1)}(1 - p_j^{(t-1)}) \\ \sum_{j=1}^N x_j p_j^{(t-1)}(1 - p_j^{(t-1)}) & \sum_{j=1}^N x_j^2 p_j^{(t-1)}(1 - p_j^{(t-1)}) \end{pmatrix}^{-1} \begin{pmatrix} \sum_{j=1}^N (y_j - p_j^{(t-1)}) \\ \sum_{j=1}^N x_j (y_j - p_j^{(t-1)}) \end{pmatrix}$$

con  $p_j^{(t-1)}$  la probabilidad estimada en la iteración  $t - 1$  calculada a partir de las estimaciones de los parámetros en la iteración  $t - 1$  de la forma

$$p_j^{(t-1)} = \frac{\exp(\beta_0^{(t-1)} + \beta_1^{(t-1)} x_j)}{1 + \exp(\beta_0^{(t-1)} + \beta_1^{(t-1)} x_j)}, \quad j = 1, \dots, N$$

Finalmente la estimación máximo verosimil de  $p_j$  viene dada por:

$$\hat{p}_j = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_j)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_j)}, \quad j = 1, \dots, N$$



siendo  $\hat{\beta}_0$  y  $\hat{\beta}_1$  los estimadores MV de los parámetros. Ésta estimación  $\hat{p}_j$  se corresponde con la estimación de la respuesta,  $\hat{y}_j$  en el caso binario.

### 2.2.2. Estimación del modelo simple binomial

De nuevo, en el modelo de regresión logística simple los dos parámetros desconocidos  $\beta_0$  y  $\beta_1$  son estimados usando el método de máxima verosimilitud. La función de verosimilitud para el escenario binomial (ver Collett [1991]) es:

$$L(\beta_0, \beta_1) = \prod_{j=1}^J \binom{n_j}{y_j} p_j^{y_j} (1 - p_j)^{n_j - y_j}$$

Tomando logaritmos se tiene:

$$\begin{aligned} \log(L(\beta_0, \beta_1)) &= \sum_{j=1}^J \left\{ \log \binom{n_j}{y_j} + y_j \log(p_j) + (n_j - y_j) \log(1 - p_j) \right\} \\ &= \sum_{j=1}^J \left\{ \log \binom{n_j}{y_j} + y_j \log \left( \frac{p_j}{1 - p_j} \right) + n_j \log(1 - p_j) \right\} \\ &= \sum_{j=1}^J \left\{ \log \binom{n_j}{y_j} + y_j (\beta_0 + \beta_1 x_j) - n_j \log(1 + \exp(\beta_0 + \beta_1 x_j)) \right\}. \end{aligned}$$

De nuevo los estimadores MV de este modelo logit siempre existen y son únicos debido a la concavidad de la log-verosimilitud (salvo separación completa). Las ecuaciones de verosimilitud no son lineales en los parámetros, de aquí que requieran métodos iterativos como el de Newton-Raphson para su resolución. Así, partiendo de las ecuaciones de verosimilitudes siguientes

$$\sum_{j=1}^J y_j x_j - \sum_{j=1}^J n_j p_j x_j = 0,$$

y resolviendo éstas por el método de Newton-Raphson se llega a la ecuación de estimación del método:

$$\beta^{(t)} = \beta^{(t-1)} + [X' \text{Diag}[n_j p_j^{(t-1)} (1 - p_j^{(t-1)})] X]^{-1} X' (y - m^{(t-1)}),$$

siendo  $m^{(t-1)} = n_j p_j^{(t-1)}$  y  $p_j^{(t-1)}$  la indicada anteriormente.

De nuevo la estimación máximo verosimil de  $p_j$  vendrá dada por:

$$\hat{p}_j = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_j)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_j)}, \quad j = 1, \dots, N$$

siendo  $\hat{\beta}_0$  y  $\hat{\beta}_1$  los estimadores MV de los parámetros, y a partir de ella la estimación de la respuesta  $\hat{y}_j = n_j \hat{p}_j$  en el caso binomial.





## Capítulo 3

# Bondad de ajuste en regresión logística

Una vez construido el modelo de regresión logística simple, tiene sentido comprobar cómo de bueno es el ajuste de los valores predichos por el modelo a los valores observados. Existen diversas formas de medir la bondad de ajuste de un modelo de regresión logística. De forma global, ésta puede ser evaluada a través de medidas tipo  $R^2$ , de la tasa de clasificaciones correctas o a través de una serie de test estadísticos. En el presente trabajo estudiaremos varios de estos test estadísticos de bondad de ajuste, conociendo sus limitaciones y poniendo de relieve sus ventajas e inconvenientes.

Siguiendo la notación expuesta en el capítulo anterior, en un test global de bondad de ajuste se contrasta la hipótesis nula:

$$H_0 : p_j = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \quad \forall j = 1, \dots, J$$

frente a la hipótesis alternativa:

$$H_1 : p_j \neq \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \quad \text{para algún } j$$

En las siguientes secciones se introducen varios test, agrupados según estén basados en los patrones de las covariables, en las probabilidades estimadas por el modelo o en residuos suavizados.

### 3.1. Test basados en patrones de las covariables

En regresión logística existen varias medidas de ajuste global para comparar la diferencia entre valores predichos y valores observados. Dos de las más populares, dada su disponibilidad en los distintos software, son el test basado en la devianza  $D$  y el estadístico  $\chi^2$  de Pearson.

#### 3.1.1. Estadístico basado en la Devianza $D$

Consideremos la función de verosimilitud para el escenario de datos agrupados en una regresión logística (simple o múltiple) se puede ver en el trabajo de Collet [Collett [1991]]:

$$L(\beta_0, \beta_1) = \prod_{j=1}^J \binom{n_j}{y_j} p_j^{y_j} (1 - p_j)^{n_j - y_j}$$

y su log-verosimilitud

$$\log L(\beta_0, \beta_1) = \sum_{j=1}^J \left\{ \log \binom{n_j}{y_j} + y_j \log p_j + (n_j - y_j) \log(1 - p_j) \right\}.$$

Sea  $\hat{L}_C = L(\hat{\beta}_0, \hat{\beta}_1)$  con  $\hat{\beta}_0, \hat{\beta}_1$  los estimadores MV de los parámetros. Bajo el modelo ajustado, la verosimilitud resulta:

$$\log \hat{L}_C = \sum_{j=1}^J \left\{ \log \binom{n_j}{y_j} + y_j \log \hat{p}_j + (n_j - y_j) \log(1 - \hat{p}_j) \right\}$$

siendo  $\hat{p}_j = \frac{\hat{y}_j}{n_j}$  la probabilidad estimada de respuesta  $Y = 1$  para el  $j$ -ésimo patrón de covariables.

El modelo saturado es aquel modelo que se ajusta perfectamente a los datos, es decir, las frecuencias de respuesta  $Y = 1$  estimadas por el modelo coinciden con las observadas, y tiene tantos parámetros libres/desconocidos como observaciones diferentes de las variables explicativas. Denotemos por  $\hat{L}_F$  la verosimilitud de este modelo, su log-verosimilitud vendrá dado por:

$$\log \hat{L}_F = \sum_{j=1}^J \left\{ \log \binom{n_j}{y_j} + y_j \log \tilde{p}_j + (n_j - y_j) \log (1 - \tilde{p}_j) \right\}.$$

Para el modelo saturado se tiene que  $\tilde{p}_j = \frac{y_j}{n_j}$ , la proporción observada de respuesta  $Y = 1$  para el  $j$ -ésimo patrón de covariable/s.

La comparación entre las dos log-verosimilitudes anteriores puede ser utilizada para medir la bondad del ajuste del modelo a los datos observados, aunque es más útil compararlas multiplicando por -2 tal diferencia, cuyo resultado es lo que se conoce como devianza o estadístico de Wilks:

$$D = -2 \log \left( \frac{\hat{L}_C}{\hat{L}_F} \right) = -2 (\log \hat{L}_C - \log \hat{L}_F)$$

De la expresión anterior, se deduce:

$$\begin{aligned} D &= 2 \sum_{j=1}^J \left( \log \left( \frac{\tilde{p}_j}{\hat{p}_j} \right) + (n_j - y_j) \log \left( \frac{1 - \tilde{p}_j}{1 - \hat{p}_j} \right) \right) \\ &= 2 \sum_{j=1}^J \left( y_j \log \left( \frac{y_j}{\hat{y}_j} \right) + (n_j - y_j) \log \left( \frac{n_j - y_j}{n_j - \hat{y}_j} \right) \right) \end{aligned}$$

y así se compara, para cada patrón de covariables, los valores observados  $y_j$  con los valores ajustados  $\hat{y}_j$ .

El estadístico así construido tiene distribución asintótica Chi Cuadrado, con grados de libertad la diferencia entre la dimensión del espacio paramétrico

## 22 CAPÍTULO 3. BONDAD DE AJUSTE EN REGRESIÓN LOGÍSTICA

y la dimensión de este espacio bajo la hipótesis nula. Así, la hipótesis nula será rechazada para el nivel de significación  $\alpha$  cuando  $D \geq \chi^2_{J-(R+1);\alpha}$  (para el caso múltiple de  $R$  covariables), que es equivalente a que el p-valor del contraste sea menor que el nivel  $\alpha$  fijado. El test así definido coincide con el test de razón de verosimilitudes para comparar el modelo saturado con el logístico binario.

La devianza puede expresarse como una suma de los cuadrados de lo que se conoce como residuos de la devianza que fueron definidos por Hosmer y Lemeshow (Hosmer et al. [1997]) de la siguiente forma.

$$D = \sum_{j=1}^J d_j^2$$

siendo

$$d_j = \text{signo}(y_j - \hat{y}_j) \left[ 2(y_j \log \left( \frac{y_j}{\hat{y}_j} \right) + (n_j - y_j) \log \left( \frac{n_j - y_j}{n_j - \hat{y}_j} \right) \right]^{1/2}$$

En el escenario binario o Bernoulli de datos no agrupados, la verosimilitud bajo el modelo ajustado es:

$$\log \hat{L}_C = \sum_{j=1}^N \{y_j \log \hat{p}_j + (1 - y_j) \log (1 - \hat{p}_j)\}$$

Bajo el modelo saturado se tiene que tanto  $y_j \log y_j$  como  $(1 - y_j) \log (1 - y_j)$  son nulos ya que  $\tilde{p}_j = y_j \in \{0, 1\}$ . Por tanto, la devianza se reduce a (Collett [1991]):

$$D = -2 \sum_{j=1}^N \left\{ \hat{p}_j \log \left( \frac{\hat{p}_j}{1 - \hat{p}_j} \right) + \log (1 - \hat{p}_j) \right\}$$

que en este caso ya no compara valores observados y ajustados, por lo que este método no puede usarse para medir la bondad del ajuste en este escenario.



### 3.1.2. Estadístico Chi Cuadrado de Pearson $\chi^2$

El estadístico Chi Cuadrado de Pearson, que compara frecuencias observadas y esperadas en un escenario binomial, se define como sigue:

$$\chi^2 = \sum_{j=1}^J \frac{(y_j - n_j \hat{p}_j)^2}{n_j \hat{p}_j (1 - \hat{p}_j)} = \sum_{j=1}^J \frac{n_j (y_j - \hat{y}_j)^2}{\hat{y}_j (n_j - \hat{y}_j)}$$

Tiene la misma distribución asintótica que la devianza, es decir, una chi cuadrado con los mismos grados de libertad. Con lo cual, la hipótesis nula será rechazada para el nivel de significación  $\alpha$  cuando  $\chi^2 \geq \chi_{J-(R+1);\alpha}^2$  (para el modelo múltiple con  $R$  covariables), que es equivalente a que el p-valor del contraste sea menor que el nivel  $\alpha$  fijado.

Este estadístico anterior también puede calcularse como la suma de los cuadrados

$$\chi^2 = \sum_{j=1}^J r_j^2.$$

con

$$r_j = \frac{y_j - n_j \hat{p}_j}{\sqrt{n_j \hat{p}_j (1 - \hat{p}_j)}}$$

que fueron denominados por Hosmer como residuos de Pearson.

Tanto para poder aplicar el test basado en la devianza como para el estadístico  $\chi^2$  tiene que verificarse que el número de observaciones para cada combinación de las variables explicativas sea grande, es por ello, por lo que estos métodos no se aplican en el caso de covariables continuas o modelos no agrupados de Bernoulli, siendo más habituales para estos casos los test desarrollados a continuación. En cuanto a las ventajas en el uso de  $D$  y  $X^2$  destaca su implementación en todos los programas estadísticos dada la simplicidad en el cálculo, mostrándose tanto el valor del estadístico como el pvalor asociado. Aunque en la mayoría de las ocasiones el valor de los dos estadísticos no es el mismo, cuando la diferencia entre ellos es grande se debe

revisar con cuidado la adecuación a la aproximación chi cuadrado, ya que suele ser indicativo que ésta no es satisfactoria (Collett [1991]).

En general, el estadístico  $D$  suele ser preferido al  $\chi^2$ , ya que es utilizado en la comparación de modelos anidados (Collett [1991]) mientras que el  $\chi^2$  no. Otra razón por la que es preferido es cuando la estimación del modelo se hace a través del método MV, porque las estimación MV de las probabilidades de éxito maximizan la función de verosimilitud para el modelo ajustado, y la devianza se ve minimizada por dichas estimaciones (Collett [1991]).

### **3.2. Test basados en probabilidades estimadas**

Hosmer y Lemeshow desarrollaron una serie de test estadísticos para medir la bondad de ajuste basados en la agrupación de las observaciones según las probabilidades estimadas por el modelo. Los dos estadísticos de este tipo más utilizados por Hosmer y Lemeshow fueron los denominados  $C_g$  y  $H_g$  cuya diferencia fundamental entre uno y otro es la forma de agrupar las probabilidades estimadas. Entre estos dos, está más extendido el uso del  $C_g$  ya que está incluido en la salida de la mayoría de los programas estadísticos. Los otros estadísticos propuestos por H-L requieren hipótesis relacionadas con el análisis discriminante, de su estimación o de integración numérica. Por no tener relación directa con la regresión logística no serán tratados aquí.

La ventaja de estos test respecto al test  $\chi^2$  y el basado en la devianza, es que se pueden utilizar tanto para datos no agrupados (modelo bernouilli) como agrupados (modelo binomial) aunque esta última situación puede influir en la formación de los grupos de probabilidades predichas. Una de las desventajas más importantes del agrupamiento de los datos según las proba-

bilidades predichas es que las desviaciones del modelo debido a un número pequeño de observaciones podrán pasar desapercibidas.

### 3.2.1. Estadístico de Hosmer-Lemeshow $C_g$

El estadístico  $C_g$  se basa en la agrupación de las probabilidades estimadas bajo el modelo de regresión  $\hat{p}_1, \dots, \hat{p}_N$ . La idea básica es que el primer grupo estará formado aproximadamente por las  $\frac{N}{G}$  observaciones cuyas probabilidades predichas sean más pequeñas, el segundo por los siguientes  $\frac{N}{G}$  más pequeños y así sucesivamente. Los puntos de corte así generados se denominan deciles de riesgo. La siguiente tabla muestra las frecuencias esperadas y observada, en cada uno de los grupos, utilizados en el cálculo del estadístico  $C_g$ , denotando por  $d_i$ ,  $i = 1, \dots, 10$ , los deciles de riesgo de las probabilidades estimadas.

Cuadro 3.1: Frecuencias esperadas y observadas para  $C_g$

	Respuesta			
	$Y = 1$		$Y = 0$	
Grupos	Observado	Esperado	Observado	Esperado
$\hat{p}_j < d_1$	$o_{11}$	$e_{11}$	$o_{01}$	$e_{01}$
$d_1 \leq \hat{p}_j < d_2$	$o_{12}$	$e_{12}$	$o_{02}$	$e_{02}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$d_9 \leq \hat{p}_j < d_{10}$	$o_{1G}$	$e_{1G}$	$o_{0G}$	$e_{0G}$
Total	$o_1$	$e_1$	$o_0$	$e_0$

El número de individuos observados para los que ocurrió el suceso y para los que no ocurrió, en cada uno de los grupos es (frecuencias observadas):

$$o_{1g} = \sum_{k=1}^{n_g} y_k$$

$$o_{0g} = \sum_{k=1}^{n_g} (1 - y_k)$$

siendo  $n_g$  el número de observaciones en el grupo  $g$ .

Análogamente, el número esperado de individuos para los que ocurrirá el suceso y para los que no, se denotan por (frecuencias esperadas):

$$e_{1g} = \sum_{k=1}^{n_g} \hat{p}(x_k)$$

$$e_{0g} = \sum_{k=1}^{n_g} (1 - \hat{p}(x_k))$$

El estadístico  $C_g$  se obtiene entonces comparando estos valores observados y esperados de la siguiente forma:

$$C_g = \sum_{k=0}^1 \sum_{g=1}^G \frac{(o_{kg} - e_{kg})^2}{e_{kg}}$$

A través de estudios de simulación se demostró que cuando  $J = N$ , si  $R + 1 < G$  (el número de covariables más 1 es menor que el número de grupos), bajo la hipótesis del modelo logístico,  $C_g$  tiene distribución asintótica  $\chi^2_{G-2}$  (Hosmer and Lemeshow [1989]).

El inconveniente en el uso de este estadístico radica en su dependencia en la elección de los puntos de corte, dando lugar a estimaciones diferentes para los mismos datos por distintos programas, pudiendo llegar incluso a darse la situación extrema de aceptación de la hipótesis nula de un ajuste adecuado por parte de algún programa y de rechazo por otro. Es por esta razón, por la que el estadístico  $C_g$  se considera algo inestable, aunque la mayor parte de los software siguen apostando por la implementación de este test.

### 3.2.2. Estadístico de Hosmer-Lemeshow $H_g$

El siguiente test propuesto por Hosmer y Lemeshow se basa en la formación de los grupos de acuerdo a unos puntos de corte fijos y preestablecidos.

El número de grupos a utilizar puede ser arbitrario, aunque los autores recomendaron el uso de 10 (Hosmer and Lemeshow [1980]). La tabla 3.2 muestra las frecuencias esperadas y observadas para cada uno de estos grupos.

Cuadro 3.2: Frecuencias esperadas y observadas para  $H_g$ 

	Respuesta			
	$Y = 1$		$Y = 0$	
Grupos	Observado	Esperado	Observado	Esperado
$0 \leq \hat{p}_j < 0,1$	$o_{11}$	$e_{11}$	$o_{01}$	$e_{01}$
$0,1 \leq \hat{p}_j < 0,2$	$o_{12}$	$e_{12}$	$o_{02}$	$e_{02}$
...	...	...	...	...
$0,9 \leq \hat{p}_j < 1,0$	$o_{110}$	$e_{110}$	$o_{010}$	$e_{010}$
Total	$o_1$	$e_1$	$o_0$	$e_0$

La formulación del estadístico y su distribución asintótica es la misma que para el anterior  $C_g$ :

$$H_g = \sum_{k=0}^1 \sum_{g=1}^G \frac{(o'_{kg} - e'_{kg})^2}{e'_{kg}}$$

En el trabajo realizado por Hosmer y Lemeshow pusieron de manifiesto que entre todos los test basados en las probabilidades estimadas que habían desarrollado, los más razonables eran estos dos,  $C_g$  y  $H_g$ . Aunque no queda claro cuál sería más adecuado entre ellos: en un primer momento, tras varias simulaciones, vieron como  $H_g$  parecía más potente que  $C_g$ , pero más tarde señalaron como más adecuado a  $C_g$  pensando que se ajustaba mejor a la distribución chi-cuadrado (Hosmer and Lemeshow [1989]). En su último trabajo sobre este tema (Hosmer et al. [1997]) los autores recomendaron usar estos estadísticos para confirmar la falta de ajuste señalada tras la utilización de otros métodos.

Nótese finalmente que estos estadísticos no son más que estadísticos chi-cuadrado de bondad de ajuste por lo que para que su distribución asintótica sea Chi Cuadrado, dicha aproximación será válida siempre que al menos el 80 % de las frecuencias estimadas bajo el modelo sean mayores que 5 y todas mayores que 1.

### 3.3. Test Score

Existen varios test score propuestos con el fin de determinar la bondad de un modelo de regresión logística. En este trabajo se introducen los desarrollados por Brown (Brown [1982]), Stukel (Stukel [1988]) y Tsiatis (Tsiatis [1980]).

#### 3.3.1. Estadístico Score de Brown $\hat{B}$

Brown desarrolló un test score que básicamente compara dos modelos. De forma general, podemos decir que se basa en considerar a los modelos logit dentro de una familia paramétrica más general de modelos (Prentice [1976]) de los que el modelo logístico resultan al particularizar ciertos valores de los parámetros (Brown [1982]). La familia general de modelos a los que hacía referencia son definidos por:

$$P_B(x_j) = \frac{\int_0^{p_j} z^{m_1-1}(1-z)^{m_2-1} dz}{B(m_1, m_2)}$$

siendo  $B(m_1, m_2)$  la función beta:

$$B(m_1, m_2) = \int_0^1 u^{m_1-1}(1-u)^{m_2-1} du,$$

$x_j$  la  $j$ -ésima observación de la variable explicativa y  $p_j = \exp(\beta_0 + \beta_1 x_j) / (1 + \exp(\beta_0 + \beta_1 x_j))$

Cuando  $m_1 = m_2 = 1$  se tiene que  $P_B(x_j) = p_j$ , y es así como se obtiene el modelo de regresión logística. Así, el test estadístico en este caso para medir la bondad del modelo contrasta la hipótesis nula  $H_0 : m_1 = m_2 = 1$ .

Los estadísticos score de los parámetros son las derivadas parciales de la log-verosimilitud  $L$  del modelo logístico con datos no agrupados con respecto a cada parámetro estimado, sustituyendo  $p_j$  por  $P_B(x_j)$ . Así, en el caso de una regresión logística con una sola variable predictora y datos no agrupados, los estadísticos score se definen como sigue:

$$\begin{aligned}\frac{\partial \ln L}{\partial \beta_0} &= \sum_{j=1}^N (y_j - P_B(x_j)) = t_0 \\ \frac{\partial \ln L}{\partial \beta_1} &= \sum_{j=1}^N x_j (y_j - P_B(x_j)) = t_1 \\ \frac{\partial \ln L}{\partial m_1} &= \sum_{j=1}^N (y_j - P_B(x_j)) \left( 1 + \frac{\log P_B(x_j)}{1 - P_B(x_j)} \right) = s_1 \\ \frac{\partial \ln L}{\partial m_2} &= - \sum_{j=1}^N (y_j - P_B(x_j)) \left( 1 + \frac{\log(1 - P_B(x_j))}{P_B(x_j)} \right) = s_2\end{aligned}$$

Así,  $s = (s_1, s_2)$  constituyen la base del ajuste propuesto por Brown, y son asintóticamente normales. El test utilizado para contrastar la hipótesis  $H_0 : m_1 = m_2 = 1$  se basa en el estadístico  $\hat{B} = s' C^{-1} s$  siendo  $C$  la matriz de covarianzas estimada de  $s$ :

$$C = \Sigma_{ss} - \Sigma_{st} \Sigma_{tt}^{-1} \Sigma_{ts}$$

con

$$\Sigma_{ts} = \Sigma'_{st} \quad \text{matriz de covarianzas de } t \text{ y } s$$

$$\Sigma_{ss} \quad \text{matriz de covarianzas de } s$$

$$\Sigma_{tt} \quad \text{matriz de covarianzas de } t$$

$$t' = (t_0, t_1)$$

En el caso de la regresión logística y sustituyendo las  $P_B(x_j)$  por  $p_j$  simple se tienen las siguientes ecuaciones:

$$\begin{aligned}\Sigma_{tt} &= \begin{pmatrix} \sum_{j=1}^N p_j q_j & \sum_{j=1}^N x_j p_j q_j \\ \sum_{j=1}^N x_j p_j q_j & \sum_{j=1}^N x_j^2 p_j q_j \end{pmatrix} \\ \Sigma_{ts} &= \begin{pmatrix} \sum_{j=1}^N p_j q_j \left(1 + \frac{\log(p_j)}{q_j}\right) & -\sum_{j=1}^N p_j q_j \left(1 + \frac{\log(q_j)}{p_j}\right) \\ \sum_{j=1}^N x_j p_j q_j \left(1 + \frac{\log(p_j)}{q_j}\right) \left(1 + \log \frac{p_j}{q_j}\right) & -\sum_{j=1}^N x_j p_j q_j \left(1 + \frac{\log(q_j)}{p_j}\right) \end{pmatrix} \\ \Sigma_{ss} &= \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} \\ s_{11} &= \sum_{j=1}^N p_j q_j \left(1 + \frac{\log(p_j)}{q_j}\right)^2 \\ s_{12} &= -\sum_{j=1}^N p_j q_j \left(1 + \frac{\log(p_j)}{q_j}\right) \left(1 + \frac{\log(q_j)}{p_j}\right) = s_{21} \\ s_{22} &= \sum_{j=1}^N p_j q_j \left(1 + \frac{\log(q_j)}{p_j}\right)^2\end{aligned}$$

siendo  $q_j = 1 - p_j$ .

Brown, a través de estudios de simulación utilizando tamaños de muestra 50, 100 y 200, mostró que  $\hat{B}$  tiene una distribución chi cuadrado con 2 grados de libertad. También investigó acerca de la potencia del estadístico cambiando la función de enlace en la hipótesis alternativa, llegando a concluir que considerando la distribución normal se obtenía una potencia pequeña, que mejoraba si se utilizaba la distribución de valores extremos y que la potencia más alta se obtenía para la distribución de Cauchy, a pesar de que ésta no pertenece a la familia general descrita por el autor. Respecto a la aplicación de este test, Brown comprobó solo la validez del método ante la existencia de covariables continuas, pero no para modelos de variables predictoras categóricas o aquellos con variables categóricas y continuas. Este método es



uno de los considerados por el programa BMDP (ver BMD) para comprobar la bondad de ajuste de un modelo de regresión logística.

### 3.3.2. Estadístico Score de Stukel $\hat{S}_{ST}$

Este estadístico es similar al propuesto por Brown y está basado también en la comparación del modelo de regresión logística con una familia más general de modelos (Stukel [1988]). Stukel propuso un modelo logístico generalizado que utiliza una función logit con dos parámetros adicionales, resultando un modelo logístico cuando ambos parámetros son nulos. El modelo adquiere la siguiente forma:

$$\log\left(\frac{p_j}{1-p_j}\right) = h_\alpha(\beta_0 + \beta_1 x_j) = g(x_j)$$

siendo  $h_\alpha(\beta_0 + \beta_1 x_j)$  una función estrictamente creciendo no lineal de  $(\beta_0 + \beta_1 x_j)$  dependiente de dos parámetros  $\alpha_1$  y  $\alpha_2$ .

Si  $\beta_0 + \beta_1 x_j \leq 0$ , o equivalentemente,  $p_j \leq \frac{1}{2}$ , se obtiene la siguiente expresión para  $h_\alpha$ :

$$h_\alpha = \begin{cases} \alpha_1^{-1} \exp((\alpha_1 |\beta_0 + \beta_1 x_j|) - 1) & \text{si } \alpha_1 > 0 \\ \beta_0 + \beta_1 x_j & \text{si } \alpha_1 = 0 \\ -\alpha_1^{-1} \log(1 - \alpha_1 |\beta_0 + \beta_1 x_j|) & \text{si } \alpha_1 < 0 \end{cases}$$

En caso contrario, si  $\beta_0 + \beta_1 x_j \geq 0$ , es decir, si  $p_j \geq \frac{1}{2}$  se obtiene:

$$h_\alpha = \begin{cases} -\alpha_2^{-1} \exp((\alpha_2 |\beta_0 + \beta_1 x_j|) - 1) & \text{si } \alpha_2 > 0 \\ \beta_0 + \beta_1 x_j & \text{si } \alpha_2 = 0 \\ \alpha_2^{-1} \log(1 - \alpha_2 |\beta_0 + \beta_1 x_j|) & \text{si } \alpha_2 < 0 \end{cases}$$

Por lo tanto, el modelo de regresión logística se tiene si  $\alpha_1 = \alpha_2 = 0$ , con lo que la hipótesis a contrastar en este test sería  $H_0 : \alpha_1 = \alpha_2 = 0$  y el estadístico del contraste adopta la forma:

$$\hat{S}_{ST} = s'_s V_s^{-1} s_s$$

con  $s'_s = (\frac{\partial \log L}{\partial \alpha_1}, \frac{\partial \log L}{\partial \alpha_2})$ ,  $L$  la función de verosimilitud y  $V_s$  calculada usando la matriz de información de Fisher como el estadístico  $C$  de Brown. Stukel demostró que el estadístico así construido tiene una distribución asintótica Chi Cuadrado con 2 grados de libertad. En el estudio de comparación realizado por Hosmer (Hosmer et al. [1997]) recomendaba el uso de este test principalmente por su mayor potencia frente a otros métodos de bondad de ajuste.

### 3.3.3. Estadístico de Tsiatis $T$

Otro estadístico, también con distribución  $\chi^2$ , fue propuesto por Tsiatis (Tsiatis [1980]) para medir la bondad de ajuste de un modelo de regresión logística. Básicamente la idea consiste en partir el espacio de covariables  $Z_1, Z_2, \dots, Z_R$  en  $k$  regiones  $R_1, R_2, \dots, R_k$ , introduciendo en el modelo un variable categórica con ese número de niveles. El modelo sería entonces:

$$\text{logit}(p_j) = \beta'Z + \gamma'I$$

con

$$I^{(j)} = \begin{cases} 1 & \text{si } (Z_1, \dots, Z_p) \in R_j \\ 0 & \text{en otro caso} \end{cases}$$

La hipótesis nula en este caso será  $H_0 : \gamma_1 = \gamma_2 = \dots = \gamma_k = 0$  y el estadístico del contraste

$$T = W'V^{-1}W$$

con una distribución asintótica  $\chi^2$  siendo  $W$  y  $V$ :

$$W = \left( \frac{\partial \log L}{\partial \gamma_1}, \dots, \frac{\partial \log L}{\partial \gamma_k} \right)$$

$$V = A - BC^{-1}B'$$

siendo  $L$  la función de verosimilitud del modelo y considerando  $A$ ,  $B$  y  $C$  como:

$$\begin{aligned} A_{jj'} &= -\frac{\partial^2 \log L}{\partial \gamma_j \partial \gamma_{j'}} \quad j, j' = 1, 2, \dots, k \\ B_{jj'} &= -\frac{\partial^2 \log L}{\partial \gamma_j \partial \beta_{j'}} \quad j = 1, 2, \dots, k \quad j' = 0, 1, \dots, m \\ C_{jj'} &= -\frac{\partial^2 \log L}{\partial \beta_j \partial \beta_{j'}} \quad j, j' = 1, 2, \dots, m \end{aligned}$$

### 3.4. Test basados en residuos suavizados

El empleo de residuos suavizados en la bondad de ajuste viene motivado por trabajos realizados en técnicas de regresión no paramétrica. La idea básica de esta serie de métodos consiste en comparar el valor suavizado de la variable respuesta para cada individuo con la estimación suavizada de las probabilidades bajo el modelo de regresión logística. En un primer momento, en el año 1983, Copas (Copas [1983]) utilizó métodos no paramétricos tipo núcleo para representar la respuesta observada y suavizada frente a la covariable. De forma análoga, Landwehr (Landwehr et al. [1984]) diseñó métodos gráficos que permitieron verificar la adecuación del modelo usando análisis cluster de vecinos próximos. En los años 1987 y 1989, Fowlkes (Fowlkes [1987]) y Azzalini (Azzalini and Härdle [1989]) continuaron el trabajo iniciado por los anteriores, desarrollando métodos basados en técnicas de suavizado.

**3.4.1. Estadístico de le Cessie y Van Houwelingen  $\hat{T}_{lc}$** 

Tanto las técnicas propuestas por los autores señalados, como la de le Cessie y van Houwelingen (le Cessie S. and Houwelingen [1991]) fueron desarrolladas para covariables de tipo continuo. El test propuesto está basado en estimaciones no paramétricas tipo núcleo de los residuos estandarizados y se utiliza fundamentalmente para datos no agrupados.

La función suavizada de los residuos es obtenida según la estimación tipo núcleo de Nadaraya y Watson (Nadaraya [1964], Watson [1964]) como una suma ponderada de los residuos de pearson  $r(x_j)$  del modelo no agrupado, definida de la siguiente forma:

$$\tilde{r}(x_j) = \frac{\sum_{k=1}^N r(x_k) K\left(\frac{x_j - x_k}{h_N}\right)}{\sum_{k=1}^N K\left(\frac{x_j - x_k}{h_N}\right)}$$

donde  $h_N$  es la ventana que controla el suavizado, que depende del tamaño de la muestra  $N$ ,  $K$  una función tipo núcleo acotada, no negativa y simétrica, normalizada según  $\int K(z)dz = 1$  y  $\int K(z)^2 dz = 1$  y los residuos de pearson

$$r_e(x_j) = \frac{y_j - \hat{p}_j}{\sqrt{\hat{p}_j(1 - \hat{p}_j)}}.$$

Para un estudio detallado de funciones de suavizado tipo núcleo en regresión, puedes consultarse los trabajos de Gasser y Müller (Gasser and Müller [1979]) y Collomb (Collomb [1981]).

El estadístico propuesto para medir la bondad del ajuste es la suma ponderada de los residuos suavizados:

$$\hat{T}_{lc} = N^{-1} \sum_{j=1}^N \tilde{r}(x_j)^2 Var(\tilde{r}(x_j))^{-1}$$

con

$$Var(\tilde{r}(x_j)) = \frac{\sum_{k=1}^N \left( K \left( \frac{x_j - x_k}{h_N} \right) \right)^2}{\left[ \sum_{k=1}^N \left( K \left( \frac{x_j - x_k}{h_N} \right) \right) \right]^2}$$

Así, se tiene que  $\hat{T}_{lc} \sim c\chi_V^2$  con  $c = \frac{Var(\hat{T}_{lc})}{2E(\hat{T}_{lc})}$  y  $V = \frac{2E(\hat{T}_{lc})^2}{Var(\hat{T}_{lc})}$  siendo  $E(\hat{T}_{lc}) = 1$  y

$$Var(\hat{T}_{lc}) = N^{-2} \sum_{j=1}^N \sum_{k=1}^N \left( \sum_{l=1}^N w_{jl}^2 \sum_{l=1}^N w_{kl}^2 \right)^{-1} \left( \sum_{l=1}^N \frac{w_{jl}^2 w_{kl}^2 (6p_l^2 - 6p_l + 1)}{p_l(1 - p_l)} + 2 \left( \sum_{l=1}^N w_{jl} w_{kl} \right)^2 \right)$$

llamando

$$w_{kl} = K \left( \frac{x_k - x_l}{h_N} \right)$$

La principal ventaja de la utilización de este estadístico radica en su no dependencia de patrones de covariables, al estar diseñado para trabajar con datos no agrupados. Otra utilidad de este test es que los elementos individuales que lo forman podrían constituir una herramienta de diagnóstico para las observaciones no ajustadas de forma correcta.

Aunque inicialmente los autores diseñaron este método para variables continuas, en un trabajo posterior (le Cessie S. and Houwelingen [1995]), le Cessie y van Houwelingen lo aplicaron a covariables categóricas y continuas. Señalaron que así como la elección de la función tipo núcleo no era trascendental, sí lo era la elección de la ventana  $h_N$ , recomendando utilizar aproximadamente  $\sqrt{N}$ . Uno de los inconvenientes en el uso del estadístico es su dificultad en cuanto al cálculo, sobre todo el relativo a su varianza. Posterior al trabajo de le Cessie y van Houwelingen, en el año 1997 Hosmer (Hosmer et al. [1997]) definió otro estadístico basado en técnicas de suavizado que queda fuera de los objetivos de este trabajo.

### 3.5. Medidas tipo $R^2$

Distintas medidas de bondad de ajuste similares al  $R^2$  usado en modelos de regresión lineal fueron propuestas para medir la bondad de ajuste de un modelo de regresión logística (Lemeshow and Hosmer [1982]). El primero de ellos fue el promedio de la proporción de variabilidad explicada (AVPE), que calcula la proporción media de la varianza de la probabilidad de un suceso. Fue denotado por Gordon (Gordon et al. [1979]) de la siguiente forma:

$$AVPE = \frac{\text{varianza incondicional de } y - \text{media varianza condicional de } y}{\text{varianza incondicional de } y}$$

considerado como varianza condicional de  $y$  a  $E[(y_j - p_j)^2 | x_j] = p_j(1 - p_j)$  y como incondicional a  $\bar{p}\bar{q}$ , siendo  $\bar{q} = 1 - \bar{p}$  y  $\bar{p} = E[p_j] = \sum_j \frac{p_j}{N}$ .

De esta forma, la proporción media de variación explicada se puede expresar como

$$AVPE = \frac{\bar{p}\bar{q} - \sum_j \frac{p_j(1 - p_j)}{N}}{\bar{p}\bar{q}}$$

El inconveniente de esta medida es que el denominador se anula en ocasiones, aunque es difícil que ésto ocurra en situaciones reales (Gordon et al. [1979]). Otro de los inconvenientes es que este estadístico no está acotado superiormente con lo que puede tomar un gran rango de valores (Gordon et al. [1979]).

Otros autores como Cox y Snell (1989), Magge (1990), y Maddala (1983) propusieron otra medida tratando de generalizar el  $R^2$  utilizado en los modelos de regresión lineal:

$$R_g^2 = 1 - \left( \frac{\hat{L}_c}{\hat{L}_0} \right)^{\frac{2}{n}}$$

siendo  $\hat{L}_c$  la logverosimilitud del modelo evaluado en  $(\hat{\beta}_0, \hat{\beta}_1)$  y  $\hat{L}_0$  la logverosimilitud del modelo que solo incluye la constante.

Por último, otra de las medidas utilizadas fue propuesta por Nagelkerke (1991) ajustando el valor máximo de  $R_g^2$  a 1:

$$\overline{R_g^2} = \frac{R_g^2}{\max(R_g^2)}$$

dónde  $\max(R_g^2) = 1 - (\hat{L}_0)^{\frac{2}{N}}$ .





# Capítulo 4

## Software

En este capítulo se detallan los métodos de bondad de ajuste disponibles en tres de los paquetes estadísticos más populares, como son R, SPSS y Stata. Se estudiará la disponibilidad en cada uno de ellos, de seis de los test estadísticos introducidos en el capítulo anterior:

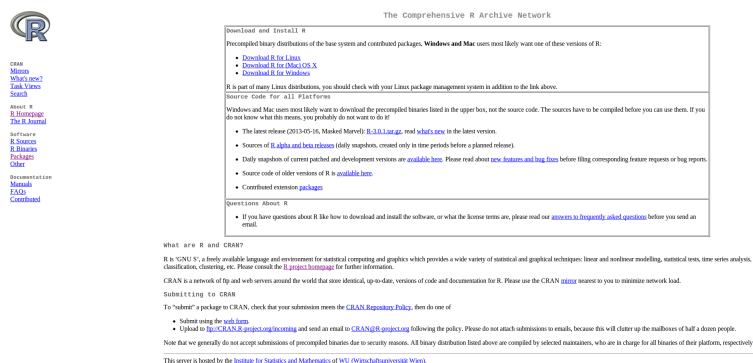
- Test basado en la Devianza  $D$
- Test basado en el estadístico  $\chi^2$
- Test Hosmer-Lemeshow  $C_g$
- Test Hommer-Lemeshow  $H_g$
- Test Score de Stukel  $\hat{S}_{ST}$
- Test de le Cessie y Van Houwelingen  $\hat{T}_{lc}$

### 4.1. R

R (consultar R) es un lenguaje de programación y un entorno para el análisis estadístico, distribuido bajo la licencia GPL de GNU (GNU), por

tanto, se trata de un software libre y gratuito. Entre sus principales características cabe destacar:

- Funciona en los sistemas operativos más habituales (Windows, Linux, Mac, etc).
- Fácil instalación.
- Dispone de numerosos complementos o paquetes para aplicaciones estadísticas concretas (actualmente están implementadas 4521 librerías o paquetes). Todas las funciones de R no definidas por el propio usuario están contenidas en dichos paquetes. Solamente cuando un paquete esté cargado estarán disponibles las definiciones que contiene.
- Está bien documentado y dispone de multitud de foros de ayuda.
- Dispone de diversas interfaces gráficas para facilitar el manejo del software, como pueden ser RCommander, Rkward o RExcel, aunque éstas son aún algo deficientes y no incluyen, por ejemplo, el cálculo de métodos de bondad de ajuste del modelo de regresión logística.
- Como todo lenguaje de programación, la curva de aprendizaje puede resultar difícil para los usuarios acostumbrados a trabajar con interfaces amigables tipo SPSS o Minitab.



The screenshot shows the CRAN website. On the left is a vertical navigation menu with links like 'CRAN', 'Mirror', 'What's new?', 'Task Views', 'Search', 'About R', 'R packages', 'The R kernel', 'Software', 'R libraries', 'R libraries', 'Packages', 'CRAN', 'Documentation', 'Manuals', 'FAQs', and 'Contact'. The main content area is titled 'The Comprehensive R Archive Network' and contains the following text:

Download and install R

Precompiled binary distributions of the base system and contributed packages, **Windows** and **Mac** users must likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for Mac OS X](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code For All Platforms

Windows and Mac users must likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2013-05-16, Marked Marvel) [R 3.0.1 for Linux](#), read [what's new](#) in the latest version.
- Sources of [R, R packages and R libraries](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current pending and developer versions are [available here](#). Please read about [new features](#) and [bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).
- Contributed extension packages

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our [guides](#), [frequently asked questions](#) before you send an email.

What are R and CRAN?

R is "GNU" R, a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc. Please consult the [CRAN homepage](#) for further information.

CRAN is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R. Please use the CRAN [mirror](#) nearest to you to minimize network load.

Submitting to CRAN

To "submit" a package to CRAN, check that your submission meets the [CRAN Repository Policy](#), then do one of

- Submit using the [web form](#)
- Upload to [ftp://CRAN.R-project.org/contrib](#) and send an email to [CRAN@R-project.org](mailto:CRAN@R-project.org) following the policy. Please do not attach submissions to emails, because this will clutter up the mailboxes of half a dozen people.

Note that we generally do not accept submissions of precompiled binaries due to security reasons. All binary distribution listed above are compiled by selected maintainers, who are in charge for all binaries of their platform, respectively.

This server is hosted by the [Institute for Statistics and Mathematics of WU \(Wirtschaftsuniversität Wien\)](#)

Para obtener información más detallada acerca de este programa pueden consultarse por ejemplo los libros de Michael J. Crawley (Crawley [2005], Crawley [2007]).

Antes de profundizar en cómo se pueden obtener los distintos métodos de bondad de ajuste del modelo con R, comenzamos revisando cómo se llevaría a cabo la construcción de un modelo de regresión logística. El ajuste de un modelo de regresión logística puede hacerse a través de la función *glm*, que no sirve solo para modelar este tipo de modelos, sino para cualquier modelo lineal generalizado. La sintaxis de esta función contenida en el paquete *stats* (ver R):

```
glm(formula, family = binomial(link="logit"), data, weights, subset,
    na.action, start = NULL, etastart, mustart, offset,
    control = list(...), model = TRUE, method = "glm.fit",
    x = FALSE, y = TRUE, contrasts = NULL, ...)
```

A continuación se describen algunos de los posibles argumentos:con:

- *formula*: describe el modelo a ajustar. En general la formulación sería  $Y = X_1 + X_2 + \dots + X_R$  siendo  $Y$  el nombre de la variable dependiente y las  $X_i$  los de las variables independientes.
- *family*: hace referencia a la familia de distribuciones y, en *link*, a la función de enlace elegida para el ajuste de este modelo. Las opciones disponibles para family son binomial, gaussian, Gamma, inverse.gaussian, poisson, quasi, quasibinomial y quasipoisson. La familia gaussian admite el link identity, log e inverse; la binomial logit, probit, cauchit, log y cloglog; la familia Gamma el link inverse, identity y log; la familia poisson el link log, identity, y sqrt; la familia inverse.gaussian el link  $\frac{1}{\mu^2}$ , inverse, identity y log.

- *data*: indica el nombre de la base de datos que contiene las variables del modelo, se trata de un argumento opcional.
- *weights*: vector opcional de pesos.
- *subset*: vector opcional que especifica el subconjunto de observaciones sobre el que quiere ajustarse el modelo.
- *na.action*: función que indica qué hacer cuando existen valores perdidos.
- *method*: indica el método utilizado para ajustar el modelo.
- *intercept*: permite controlar si se incluye en el modelo el término constante.

Parte de la salida de esta función puede obtenerse a través de la función *summary* (cuya sintaxis es *summary(nombremodelo)*, siendo *nombremodelo* el nombre del modelo ajustado). La lista siguiente es parte de la salida de la función *glm*:

- *coefficients*: vector de coeficientes.
- *residuals*: vector que contiene los residuos obtenidos en la última iteración.
- *fitted.values*: vector con los valores medios ajustados, obtenidos según la transformación de los predictores lineales por la inversa de la función de enlace.
- *family*: devuelve la familia utilizada en la construcción del modelo.
- *deviance*: devianza del modelo ajustado o -2 veces el máximo de la log verosimilitud.

- *aic*: criterio de información de Akaike, que corresponde con menos 2 veces la log verosimilitud maximizada más dos veces el número de parámetros.
- *null.deviance*: devianza para el modelo nulo que solo contiene la constante.
- *iter*: número de iteraciones.

En lo relativo a la bondad del modelo, hay algunas funciones que ya están implementadas en R y otras que hay que construir a través de sintaxis. Veamos a continuación cómo poder calcular los distintos métodos de bondad de ajuste.

- Test basado en la devianza. Se obtiene a través de la instrucción:

```
anova(modelo, test="Chisq")
```

indicando por *modelo* el nombre del modelo para el que se quiere ejecutar el contraste.

- Test basado en el estadístico  $\chi^2$ . El valor del estadístico se puede calcular como la suma de los cuadrados de los residuos de Pearson y después calcular el pvalor del contraste teniendo en cuenta que sigue la misma distribución asintótica que la devianza.

```
estadistico <- sum(residuals(modelo2, type="pearson")^2)
p <- 1 - pchisq(estadistico, gl)
```

- Test  $C_g$  y  $H_g$  de Hosmer-Lemeshow y le Cessie y Van Houwelingen. Estos tres test se encuentran disponibles en el paquete *MKmisc* (Kohl) a través de la función *HLgof.test* que tiene la siguiente sintaxis:

```
HLgof.test(fit, obs, ngr = 10, X, verbose = FALSE)
```

Los argumentos de la función son:

- *fit*: vector numérico que contiene las probabilidades estimadas por el modelo.
  - *obs*: vector numérico con valores observados.
  - *ngr*: número de grupos a utilizar en el cálculo del estadístico  $C_g$  y  $H_g$ . Hosmer y Lemeshow recomendaron el uso de 10 grupos.
  - *X*: covariable para el ajuste de le Cessie y Van Houwelingen.
  - *verbose*: valor lógico para obtener las salidas intermedias.
- Test Score de Stukel. Este test puede llevarse a cabo a través de la siguiente función (ver Presnell):

```
stukel <- function(object, alternative = c("both", "alpha1", "alpha2"))
{
  DNAME <- deparse(substitute(object))
  METHOD <- "Stukel's test of the logistic link"
  alternative <- match.arg(alternative)
  eta <- predict(object, type = "link")
  etasq <- 0.5 * eta * eta
  etapos <- eta > 0
  dv <- matrix(0, nrow = length(eta), ncol = 2)
  dv[etapos,1] <- etasq[etapos]
  dv[!etapos,2] <- - etasq[!etapos]
  colnames(dv) <- c("z1", "z2")
  oinfo <- vcov(object)
  oX <- qr.X(object$qr)
```

```

ImH <- - oX %*% oinfo %*% t(oX)
diag(ImH) <- 1 + diag(ImH)
wdv <- sqrt(object$weights) * dv
qmat <- t(wdv) %*% ImH %*% wdv
sc <- apply(dv * (object$weights * residuals(object, "working")), 2, sum)
allstat <- c(sc * sc / diag(qmat), sc %*% solve(qmat) %*% sc)
names(allstat) <- c("alpha1", "alpha2", "both")
allpar <- c(1,1,2)
names(allpar) <- names(allstat)
allpval <- pchisq(allstat, allpar, lower.tail=FALSE)
STATISTIC <- allstat[alternative]
PARAMETER <- allpar[alternative]
names(PARAMETER) <- "df"
PVAL <- allpval[alternative]
names(allpar) <- rep("df", 3)
structure(list(statistic = STATISTIC,
               parameter = PARAMETER,
               p.value = PVAL,
               alternative = alternative,
               method = METHOD, data.name = DNAME,
               allstat = allstat, allpar = allpar, allpval = allpval
            ),
          class = "htest")
}

```

Los argumentos de esta función son:

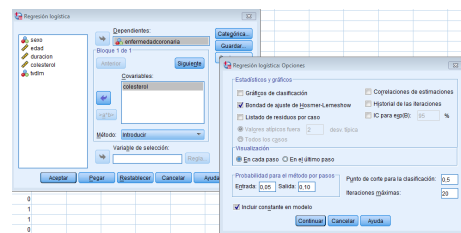
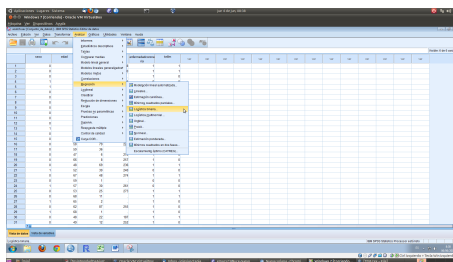
- `object`: nombre del modelo ajustado.
- `alternative`: hace referencia a la hipótesis a contrastar. Como se comentó en el capítulo anterior el modelo de regresión logística se

obtiene cuando ambos parámetros son nulos:  $\alpha_1 = \alpha_2 = 0$ .

## 4.2. SPSS

SPSS (Statistical Package for the Social Sciences) (SPS) es un programa de IBM para el análisis estadístico de datos. A través de una interfaz amigable para el usuario SPSS aborda gran variedad de técnicas y métodos estadísticos, y junto con el programa BMDP es uno de los más utilizados en Ciencias Sociales (Bausela [2005]). Permite el manejo de técnicas tal vez más específicas del ámbito biosanitario, tales como análisis de supervivencia de Kaplan Meier, modelos de regresión de Cox o modelos de regresión para respuestas discretas incluyendo la regresión logística. Para un mayor detalle de las técnicas y módulos implementados puede consultarse el centro de información disponible en la propia página de IBM (ver IBM).

En el procedimiento *Regresión Logística* de SPSS, disponible en el módulo base del programa a través del menú Analizar  $\Rightarrow$  Regresión Logística, se proporciona el ajuste y métodos para medir la bondad de un modelo de regresión logística. Por defecto aparece en la ventana de resultados el resultado del test de razón de verosimilitudes, aunque en el botón *Opciones* del mismo menú puede solicitarse también la prueba de Hosmer y Lemeshow basada en el estadístico  $C_g$ .





## 4.3. Stata

Programa para el análisis de datos bajo licencia y que pone a disposición del usuario una interfaz amigable similar a la proporcionada por SPSS. Para un mayor detalle sobre este software puede consultarse su página web (ver Sta) o los numerosos tutoriales disponibles en la red. En lo relativo al ajuste de modelos de respuesta discreta es recomendable el libro de J. Scott Long y Jeremy Freese (Long and Freese [2006]).

En lo que respecta a la construcción de un modelo de regresión logística se utiliza el siguiente comando (también accesible a través del menú de Stata):

```
logit var cov1 cov2 cov3
```

siendo *var* el nombre de la variable respuesta y *cov1*, *cov2*, *cov3* los de las covariables. En esta primera aproximación se proporciona el valor de la logverosimilitud en cada iteración, junto con una estimación de los coeficientes y su significación. Como medidas de bondad de ajuste se proporciona el pseudo  $R^2$  de McFadden junto el valor del estadístico Chi Cuadrado de razón de verosimilitudes ( *LR chi2*) y el pvalor asociado ( *Prob > chi2*).

Una vez ajustado el modelo, a través del comando *fitstat* se pueden obtener otras medidas de bondad de ajuste como son:

- Medidas tipo  $R^2$ . Se provee del  $R^2$  propuesto por McFadden, Efron, máxima verosimilitud o Maddala, McKelvey y Zavoina, y el de Cragg-Uhler.
- AIC. Criterio de información de Akaike.
- BIC. Criterio de información Bayesiana.
- Devianza. Valor de la devianza para el modelo ajustado ( $D$ ).

- Test Chi cuadrado basado en la devianza o test de razón de verosimilitudes. Se proporciona el valor del estadístico de Wilks ( $LR$ ) y el pvalor correspondiente ( $Prob > LR$ ).

Si estamos trabajando con covariables de tipo continuo y necesitamos obtener el estadístico  $C_g$  de Hosmer-Lemeshow, simplemente escribimos la siguiente sentencia y tras la presentación de la agrupación en 10 grupos basados en los deciles, se obtiene el valor del estadístico (*Hosmer-Lemeshow chi2*) y el pvalor del contraste ( $Prob > chi2$ ).

```
estat gof, table group(10)
```

Existe otra forma de plantear el modelo de regresión logística que es a través de los modelos lineales generalizados, a través de la siguiente sintaxis:

```
. binreg var cov1 cov2 cov3, or
```

siendo de nuevo *var* el nombre de la variable respuesta y *cov1*, *cov2*, *cov3* los de las covariables, *or* indica que mostrará por pantalla los OR.





# Capítulo 5

## Aplicación con datos reales

### 5.1. Descripción de los datos

Para aplicar algunos de los métodos expuestos en este trabajo, se trabajará con un conjunto de datos reales del Centro de Enfermedades Cardiovasculares de la Universidad de Duke, disponibles en la web de la Universidad de Vanderbilt: <http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets>. Se trata de una muestra de 3504 pacientes que acudieron al centro con dolor en el pecho, para los que se recogieron diversas variables cuyo nombre en la base de datos y descripción es la siguiente:

- *enf*: variable binaria que toma valores 1 y 0, indicando si el paciente presenta estrechamiento de al menos un 75 % de alguna de las arterias coronarias importante ( $enf = 1$ ) o no ( $enf = 0$ ).
- *sexo*: variable categórica que indica el sexo del paciente.
- *edad*: variable continua que representa la edad en años del individuo.
- *colesterol*: variable continua que expresa los Mg/dl de colesterol.

- *duracion*: variable continua que recoge la duración, en días, de los síntomas de la enfermedad coronaria.

Los datos fueron encontrados en formato texto, *acatch.csv*. El software elegido para hacer el análisis descriptivo fue R, luego el fichero de datos se importó para construir posteriormente la base de datos *datos* en formato *rda*.

```
datos <-read.table("/Documentos/Master UGR/TFM/acath.csv",header=TRUE,
                  sep=";", na.strings="", dec=",", strip.white=TRUE)

save(datos,file="datos.rda")
```

A continuación se muestra un resumen de cada variable recogida en la base de datos. Para las variables cualitativas se proporciona la distribución de frecuencias, mientras que para las variables cuantitativas se muestran los valores medios, mínimo, máximo y cuartiles.

La variable *sexo* es una variable categórica con 3504 casos válidos, para la que se obtiene la siguiente distribución de frecuencias:

```
> table(datos$sexo)
hombre  mujer
  2405    1099

> 100*table(datos$sexo)/sum(table(datos$sexo))
hombre  mujer
68.63584 31.36416
```

La otra variable cualitativa es *en* que será la variable dependiente, la que nos interesa modelar a través de la regresión logística, aquella que como ya se ha comentado indica si el paciente posee o no una enfermedad coronaria. Tal y como se aprecia a continuación, 2234 individuos, que corresponden al 66.61 % de la muestra, tienen indicios de tal enfermedad.

```
> table(datos$enf)
 0    1
1170 2334
> 100*table(datos$enf)/sum(table(datos$enf))
 0    1
33.39041 66.60959
```

Para las cuatro variables restantes, todas ellas continuas, se realiza el siguiente resumen numérico:

```
> summary(datos$edad)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 17.00  46.00  52.00  52.28  59.00  82.00
> summary(datos$duracion)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0     4    18    43    60    416
> summary(datos$colesterol)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 29.0   196.0   224.5   229.9   259.0   576.0   1246
```

Tal y como se aprecia, la única variable con valores perdidos, identificados por *NA* es la variable *colesterol*, que tiene 1246.

Con el fin de determinar los factores que explican la aparición de una enfermedad coronaria y dado el carácter dicotómico de esta variable, ajustaremos un modelo de regresión logística para predecir tal hecho. En un primer momento se ajustará un modelo de regresión simple, considerando como única variable predictora la variable *colesterol* y más adelante ajustaremos el modelo considerando el resto de las variables recogidas en la base de datos como variables explicativas.

## 5.2. Modelo de regresión logística simple

Comenzamos ajustando un modelo de regresión logística simple considerando como única variable explicativa la variable *colesterol*.

### 5.2.1. Ajuste y bondad del modelo con R

Antes de estimar el modelo indicado, dado que la variable *colesterol* contiene valores perdidos, utilizamos la función *drop.levels* de la librería *gdata* (Warnes [2012]) para trabajar sin ellos, pasando a trabajar a partir de ahora con la base de datos *Datos* en lugar de *datos*. Una vez dado este paso, ya podemos estimar nuestro modelo, que llamaremos *modelo1*:

```
>library(gdata)
>Datos <- drop.levels(datos[ !is.na(datos$colesterol), ])

>modelo1 <- glm(enf~colesterol,family=binomial(link=logit),data=Datos)
```

A través del comando *summary* obtenemos el resumen del modelo:

```
>summary(modelo1)

Call:
glm(formula = enf ~ colesterol, family = binomial(link = logit),
    data = Datos)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.2140	-1.3669	0.8247	0.9406	1.4279



Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.7525280	0.2186516	-3.442	0.000578 ***
colesterol	0.0062268	0.0009525	6.538	6.25e-11 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2895.3 on 2257 degrees of freedom  
 Residual deviance: 2849.7 on 2256 degrees of freedom  
 AIC: 2853.7

Number of Fisher Scoring iterations: 4

En el resumen del modelo, en primer término aparece una descripción de los residuos de la devianza, en particular, los valores mínimo, máximo, primer, segundo y tercer cuartil. Después se presenta la tabla de coeficientes, errores estándar, valor del estadístico de Wald y el pvalor asociado. Por último, la salida incluye el valor de la devianza para el modelo que solo incluye la constante (*Null deviance*) seguida de sus grados de libertad, a continuación estaría el valor de la devianza para el modelo ajustado (*Residual deviance*) y sus grados de libertad, y por último el valor *AIC*. Se indica asimismo el número de iteraciones utilizadas.

Estudiamos a continuación la bondad del modelo construido. Comenzamos realizando el análisis de la devianza tal y como se detalló en el capítulo anterior.

```
> anova(modelo1, test="Chisq")
```

Analysis of Deviance Table

Model: binomial, link: logit

Response: enf

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			2257	2895.3	
colesterol	1	45.637	2256	2849.7	1.423e-11 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

La salida de este análisis nos proporciona una tabla dónde vemos por filas ciertos resultados para el modelo nulo y para el modelo ajustado. Por columnas nos aparece en primer término los grados de libertad de la distribución asintótica de la devianza ( $Df$ ), seguido de *Deviance* que muestra la variación en la devianza al comparar los dos modelos, y a continuación *Resid.Df*, *Resid.Dev* y *Pr(>Chi)* que corresponden respectivamente a los grados de libertad de la devianza de cada modelo, el valor del estadístico y el pvalor del contraste, que en este caso es muy significativo.

Para efectuar el análisis basado en el estadístico Chi Cuadrado, calculamos el valor del estadístico como la suma de cuadrados de los residuos de Pearson, para después calcular la significación del test. Como se aprecia, el test sigue siendo significativo.

```
> sum(residuals(modelo1,type="pearson")^2)
2256.21
> 1 - pchisq(sum(residuals(modelo1,type="pearson")^2),1)
0
```

A continuación realizamos los dos estadísticos de Hosmer-Lemeshow y el de le Cessie y Van Houwelingen.

```

> HLgof.test(fit = fitted(modelo1),
  obs = as.numeric(as.character(Datos$enf)),
  X = cbind(Datos$colesterol))

$C
Hosmer-Lemeshow C statistic
data: fitted(modelo1) and as.numeric(as.character(Datos$enf))
X-squared = 9.0287, df = 8, p-value = 0.3399

$H
Hosmer-Lemeshow H statistic
data: fitted(modelo1) and as.numeric(as.character(Datos$enf))
X-squared = 19.0765, df = 8, p-value = 0.01446

$gof
le Cessie-van Houwelingen-Copas-Hosmer global goodness of fit test
data: fitted(modelo1) and as.numeric(as.character(Datos$enf))
z = 0.1927, p-value = 0.8472

```

Ayudados por la función definida en el capítulo anterior, calculamos la significación del contraste de Stukel.

```

>stukel(modelo1,alternative="both")

Stukel's test of the logistic link
data: modelo1
both = 5.9031, df = 2, p-value = 0.05226

```

A modo de resumen, los resultados obtenidos con R serían:

<b>Estadístico</b>	<b>pvalor</b>
Devianza	0.0000
Chi Cuadrado	0.0000
$C_g$	0.3399
$H_g$	0.0145
$T_{lk}$	0.8472
Stukel	0.0523

Considerado un nivel de significación  $\alpha = 0,05$ , tres de los seis procedimientos arrojan pvalores inferiores, por lo tanto resultan significativos a ese nivel. Si bien, dos de ellos, el basado en la devianza y el Chi Cuadrado, no son adecuados en este caso porque la aproximación asintótica no es válida para variables continuas. Así, obviando el resultado de estos dos, sólo el contraste basado en el estadístico  $H_g$  indica un mal ajuste para el nivel de significación señalado.

### 5.2.2. Ajuste y bondad del modelo con SPSS

Al realizar el modelo de regresión con SPSS, se nos muestra en la primera tabla un resumen de los casos analizados: hay un total de 3504 individuos pero no se consideran los 1246 perdidos en el análisis.

### ➔ Regresión logística

[Conjunto\_de\_datos1]

**Resumen del procesamiento de los casos**

Casos no ponderados <sup>a</sup>		N	Porcentaje
Casos seleccionados	Incluidos en el análisis	2258	64,4
	Casos perdidos	1246	35,6
	Total	3504	100,0
Casos no seleccionados		0	,0
Total		3504	100,0

a. Si está activada la ponderación, consulte la tabla de clasificación para ver el número total de casos.

**Codificación de la variable dependiente**

Valor original	Valor interno
0	0
1	1

A continuación se presenta el ajuste para el modelo que sólo tiene la constante, es lo que identifica por Bloque0, mostrando la tabla de clasificaciones correctas, las variables incluidas y no incluidas en el modelo considerado en este primer paso. Como variables incluidas solo está la constante, para la que proporciona su coeficiente ( $B$ ), error estándar ( $E.T$ ), valor del estadístico de Wald ( $Wald$ ) y significación ( $Sig.$ ), junto con la exponencial del coeficiente ( $exp(B)$ ). En la última de estas tablas comprobamos como no incluye la variable *colesterol* como variable predictora.

**Bloque 0: Bloque inicial****Tabla de clasificación<sup>a, b</sup>**

Observado			Pronosticado		
			enfermedadcoronaria		Porcentaje correcto
			0	1	
Paso 0	enfermedadcoronaria	0	0	768	,0
		1	0	1490	100,0
Porcentaje global					66,0

a. En el modelo se incluye una constante.

b. El valor de corte es ,500

**Variables en la ecuación**

	B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 0 Constante	,863	,044	222,593	1	,000	1,940

**Variables que no están en la ecuación**

	Puntuación	gl	Sig.
Paso 0 Variables colesterol	43,773	1	,000
Estadísticos globales	43,773	1	,000

En el siguiente paso ya se introduce el *colesterol* como variable explicativa:

**Bloque 1: Método = Introducir****Pruebas omnibus sobre los coeficientes del modelo**

	Chi cuadrado	gl	Sig.
Paso 1 Paso	45,637	1	,000
Bloque	45,637	1	,000
Modelo	45,637	1	,000

**Resumen del modelo**

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	2849,650 <sup>a</sup>	,020	,028

a. La estimación ha finalizado en el número de iteración 4 porque las estimaciones de los parámetros han cambiado en menos de ,001.

**Tabla de clasificación<sup>a</sup>**

Observado			Pronosticado		
			enfermedadcoronaria		Porcentaje correcto
			0	1	
Paso 1	enfermedadcoronaria	0	7	761	,9
		1	17	1473	98,9
Porcentaje global					65,5

a. El valor de corte es ,500

**Variables en la ecuación**

	B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1 <sup>a</sup> colesterol	,006	,001	42,740	1	,000	1,006
Constante	-,753	,219	11,845	1	,001	,471

a. Variable(s) introducida(s) en el paso 1: colesterol.

En la tabla *Prueba omnibus sobre los coeficientes del modelo* aparece el valor de menos dos veces la diferencia entre la logverosimilitud del modelo ajustado y el modelo nulo que solo incluye la constante (*Chi cuadrado*), sus grados de libertad (*gl*) y su significación (*Sig.*). La fila primera (PASO) es la correspondiente al cambio de verosimilitud (de  $-2LL$ ) entre pasos sucesivos en la construcción del modelo, contrastando la  $H_0$  de que los coeficientes de las variables añadidas en el último paso son cero. La segunda fila (BLOQUE) es el cambio entre bloques de entrada sucesivos durante la construcción del modelo. Si como es habitual en la práctica se introducen las variables en un solo bloque, el Chi Cuadrado del Bloque es el mismo que el Chi Cuadrado del Modelo. La tercera fila (MODELO) es la diferencia entre el valor de  $-2LL$  para el modelo nulo y el valor de  $-2LL$  para el modelo actual. Como en nuestro caso, solo hay una variable explicativa, un único bloque y un único paso, las tres filas son exactamente iguales.

En la tabla *Resumen del modelo* aparece el valor de la devianza ( $-2 \log \text{verosimilitud}$ ), y el  $R^2$  de Cox-Snell y de Nagelkerke. La Tabla de Clasificación especifica el % de clasificaciones correctas si se elige como punto de corte 0.5, que en nuestro caso es del 65.5 %. Y por último en la tabla *Variables en la ecuación* se obtiene la misma tabla que en el Bloque0, pero incluyendo una nueva fila para la variable introducida en el último paso.

Como en el botón correspondiente del menú de Regresión logística marcamos el test de Hosmer-Lemeshow, se muestra el valor del estadístico  $C_g$  (*Chi Cuadrado*), sus grados de libertad (*gl*) y su significación (*Sig.*). En la última parte de la salida se muestra la tabla de frecuencias de los valores esperados y observados utilizada en el cálculo del estadístico anterior.

Prueba de Hosmer y Lemeshow			
Paso	Chi cuadrado	gl	Sig.
1	8,241	8	,410

Tabla de contingencias para la prueba de Hosmer y Lemeshow					
		enfermedad coronaria = 0		enfermedad coronaria = 1	
		Observado	Esperado	Observado	Esperado
Paso 1	1	93	99,564	128	121,436
	2	96	86,511	117	126,489
	3	84	85,822	138	136,178
	4	85	83,521	141	142,479
	5	89	79,130	135	144,870
	6	78	74,344	142	145,656
	7	64	73,804	166	156,196
	8	72	69,819	161	163,181
	9	60	60,775	160	159,225
	10	47	54,711	202	194,289
					249

Así, la tabla resumen con la significación obtenida según los métodos disponibles en SPSS sería:

Estadístico	pvalor
Devianza	0.0000
$C_g$	0.410

### 5.2.3. Ajuste y bondad del modelo con Stata

Comenzamos ajustando el modelo de regresión. En la salida proporcionada aparece el logaritmo de la verosimilitud para cada iteración, a continuación una serie de medidas que se comentarán más adelante, a la vez que se comente la salida de otro comando, y por último, se muestra la tabla de coeficientes estimados, errores estándar, valor del estadístico de Wald, su significación e intervalos de confianza al 95 %.

```
. logit enf colesterol
```

```
Iteration 0: log likelihood = -1447.6438
```

```
Iteration 1: log likelihood = -1424.9581
```

```
Iteration 2: log likelihood = -1424.8251
```



Iteration 3: log likelihood = -1424.8251

Logistic regression	Number of obs	=	2258
	LR chi2(1)	=	45.64
	Prob > chi2	=	0.0000
Log likelihood = -1424.8251	Pseudo R2	=	0.0158

enf	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
cholesterol	.0062268	.0009525	6.54	0.000	.00436	.0080936
_cons	-.752528	.2186517	-3.44	0.001	-1.181077	-.3239786

Una vez construido el modelo, simplemente escribiendo la sentencia *fitstat* obtenemos un resumen con ciertas pruebas de bondad de ajuste. En primer término, las tres primeras filas de la tabla hacen referencia al análisis de la devianza, mostrando la logverosimilitud para el modelo nulo (*Log-Lik Intercept Only*), para el ajustado (*Log-Lik Full Model*), el valor y gl de la devianza ( $D(gl)=valor$ ), la variación en la devianza al comparar los dos modelos ( $LR(1)$ ) y el pvalor asociado ( $Prob > LR$ ). Estos dos últimos valores son los que junto con el logaritmo de la verosimilitud del modelo ajustado, fueron proporcionados cuando se ajustó el modelo. El resto de medidas listadas, las tipo  $R^2$  y BIC/AIC, no son objeto de estudio de este trabajo, por lo que no entraremos en detalle.

```
. fitstat
```

```
Measures of Fit for logit of enf
```

Log-Lik Intercept Only:	-1447.644	Log-Lik Full Model:	-1424.825
D(2256):	2849.650	LR(1):	45.637
		Prob > LR:	0.000
McFadden's R2:	0.016	McFadden's Adj R2:	0.014
Maximum Likelihood R2:	0.020	Cragg & Uhler's R2:	0.020
McKelvey and Zavoina's R2:	0.029	Efron's R2:	0.019
Variance of y*:	3.389	Variance of error:	3.290
Count R2:	0.655	Adj Count R2:	-0.013
AIC:	1.264	AIC*n:	2853.650
BIC:	-14571.711	BIC':	-37.915

A continuación realizamos el test de Hosmer Lemeshow, obteniendo en primer lugar la tabla de frecuencias observadas y esperadas para cada valor de la respuesta, en cada uno de los 10 grupos elegidos. Después se indica el número de observaciones, número de grupos, el valor del estadístico  $C_g$  (*Hosmer-Lemeshow  $\chi^2(gl)$* ) y el pvalor del contraste ( $Prob > \chi^2$ ).

```
. estat gof, table group(10)
Logistic model for enf, goodness-of-fit test
(Table collapsed on quantiles of estimated probabilities)
+-----+
| Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
|-----+-----+-----+-----+-----+-----+-----|
|      1 | 0.5835 |   141 | 134.3 |   102 | 108.7 |   243 |
|      2 | 0.6060 |   127 | 136.1 |   101 |  91.9 |   228 |
|      3 | 0.6237 |   134 | 131.2 |    79 |  81.8 |   213 |
|      4 | 0.6411 |   145 | 147.4 |    88 |  85.6 |   233 |
|      5 | 0.6560 |   127 | 137.5 |    85 |  74.5 |   212 |
|-----+-----+-----+-----+-----+-----+-----|
|      6 | 0.6719 |   148 | 155.4 |    86 |  78.6 |   234 |
```

```

|      7 | 0.6909 |   171 | 160.4 |    64 |  74.6 |   235 |
|      8 | 0.7130 |   152 | 151.3 |    63 |  63.7 |   215 |
|      9 | 0.7450 |   165 | 160.7 |    56 |  60.3 |   221 |
|     10 | 0.9445 |   180 | 175.7 |    44 |  48.3 |   224 |

```

```

+-----+

```

```

      number of observations =      2258
      number of groups      =         10
Hosmer-Lemeshow chi2(8)    =         9.00
      Prob > chi2           =      0.3419

```

Al plantear el modelo dentro de la clase de modelos lineales generalizado, podemos obtener el valor del estadístico Chi Cuadrado (*Pearson*), además del de la devianza ya obtenida.

```

. binreg enf colesterol, or

```

```

Iteration 1:  deviance =  2854.368
Iteration 2:  deviance =  2849.652
Iteration 3:  deviance =   2849.65
Iteration 4:  deviance =   2849.65

```

```

Generalized linear models          No. of obs      =      2258
Optimization      : MQL Fisher scoring      Residual df      =      2256
                  (IRLS EIM)                Scale parameter =         1
Deviance          =  2849.650106             (1/df) Deviance =  1.263143
Pearson           =  2256.209507             (1/df) Pearson  =  1.000093

```

```

Variance function: V(u) = u*(1-u)          [Bernoulli]
Link function      : g(u) = ln(u/(1-u))     [Logit]

```

BIC = -14571.71

-----						
		EIM				
enf		Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
-----+						
cholesterol		1.006246	.0009584	6.54	0.000	1.00437 1.008126
-----						

A partir de esta salida ya podríamos calcular la significación del contraste:

```
. display "pvalueChi2=" chiprob(1, 2256.209507)
pvalueChi2=0
```

Luego, la tabla resumen con la significación obtenida según los métodos disponibles en Stata sería:

Estadístico	pvalor
Devianza	0.0000
Chi Cuadrado	0.0000
$C_g$	0.3419

### 5.3. Modelo de regresión logística múltiple

Realizamos a continuación el ajuste del modelo considerando ahora todas las variables recogidas como variables explicativas, como son: el sexo, la edad, la duración del dolor en el pecho y el colesterol.

#### 5.3.1. Ajuste y bondad del modelo con R

Estimamos el modelo completo, el cual identificaremos por *modelo2*.

```

> modelo2 <- glm(enf~colesterol+duracion+sexo+edad,family=binomial(link=logit),
                  data=Datos)

> summary(modelo2)

Call:
glm(formula = enf ~ colesterol + duracion + sexo + edad,
     family = binomial(link = logit), data = Datos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5401  -0.8704   0.5282   0.7663   2.4125

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.286432    0.386800  -11.082  <2e-16 ***
colesterol    0.009123    0.001079   8.453  <2e-16 ***
duracion     -0.001695    0.001014  -1.672  0.0945 .
sexomujer    -2.101717    0.113559 -18.508  <2e-16 ***
edad          0.073032    0.006146  11.883  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2895.3  on 2257  degrees of freedom
Residual deviance: 2344.6  on 2253  degrees of freedom
AIC: 2354.6
Number of Fisher Scoring iterations: 4

```

Calculamos el estadístico Devianza, el Chi Cuadrado y la significación de los contrastes:

```

> sum(residuals(modelo2,type="deviance")^2)

2344.626

```

```

> 1 - pchisq(sum(residuals(modelo2,type="deviance")^2),1)
0

> sum(residuals(modelo2,type="pearson")^2)
2275.631

> 1 - pchisq(sum(residuals(modelo2,type="pearson")^2),1)
0

```

Aplicamos a continuación la función *HLgof.test* al modelo ajustado para obtener los dos estadísticos de Hosmer y Lemeshow, junto el de le Cessie y Van Houwelingen.

```

> HLgof.test(fit = fitted(modelo2), obs = as.numeric(as.character(enf)),
             X= model.matrix(enf ~ colesterol+edad+sexo+duracion))

$C
Hosmer-Lemeshow C statistic
data:  fitted(modelo2) and as.numeric(as.character(enf))
X-squared = 7.0671, df = 8, p-value = 0.5294

$H
Hosmer-Lemeshow H statistic
data:  fitted(modelo2) and as.numeric(as.character(enf))
X-squared = 21.6106, df = 8, p-value = 0.005691

$gof
le Cessie-van Houwelingen-Copas-Hosmer global goodness of fit test
data:  fitted(modelo2) and as.numeric(as.character(enf))
z = -0.5035, p-value = 0.6146

```

Y por último, el resultado del test Score de Stukel sería:

```

> stukel(modelo2)

```

```

Stukel's test of the logistic link
data:  modelo2
both = 12.1978, df = 2, p-value = 0.002245
alternative hypothesis: both

```

En este caso, de nuevo, los resultados de los test Chi Cuadrado y el de la Devianza no serían válidos. Los resultados de los otros cuatro no son concordantes entre sí, ya que dos de ellos aceptan el buen ajuste a los datos ( $C_g$  y  $T_{lk}$ ) mientras que los otros dos lo rechazan ( $H_G$  y Stukel).

Estadístico	pvalor
Devianza	0.0000
Chi Cuadrado	0.0000
$C_g$	0.5294
$H_g$	0.0057
$T_{lk}$	0.6146
Stukel	0.0022

### 5.3.2. Ajuste y bondad del modelo con SPSS

De forma análoga a lo explicado para la regresión simple, construimos el modelo incluyendo todas las variables. Se muestran los resultados de los dos métodos de bondad de ajuste proporcionados por SPSS, obteniendo resultados en la misma dirección que R, si bien el pvalor del test  $C_g$  es ligeramente distinto (con SPSS se obtiene 0.416 y con R 0.5294). (pvalor=0.000) mientras que el test de Hosmer y Lemeshow la aceptaría (pvalor=0.416).

**Bloque 1: Método = Introducir****Pruebas omnibus sobre los coeficientes del modelo**

		Chi cuadrado	gl	Sig.
Paso 1	Paso	550,661	4	,000
	Bloque	550,661	4	,000
	Modelo	550,661	4	,000

**Resumen del modelo**

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	2344,626 <sup>a</sup>	,216	,299

a. La estimación ha finalizado en el número de iteración 5 porque las estimaciones de los parámetros han cambiado en menos de ,001.

**Prueba de Hosmer y Lemeshow**

Paso	Chi cuadrado	gl	Sig.
1	7,719	8	,461

**Tabla de contingencias para la prueba de Hosmer y Lemeshow**

		enfermedad coronaria = 0		enfermedad coronaria = 1		Total
		Observado	Esperado	Observado	Esperado	
Paso 1	1	174	182,259	52	43,741	226
	2	151	143,905	75	82,095	226
	3	115	112,060	111	113,940	226
	4	93	85,269	133	140,731	226
	5	69	67,347	157	158,653	226
	6	59	54,558	167	171,442	226
	7	35	43,655	191	182,345	226
	8	33	34,939	193	191,061	226
	9	23	27,087	203	198,913	226
	10	16	16,922	208	207,078	224

**5.3.3. Ajuste y bondad del modelo con Stata**

Respecto al ajuste del modelo con Stata, comenzamos estimándolo a través de la orden *logit*, que además de la estimación de los coeficientes ya comentada, indica el cambio en la log verosimilitud al comparar el modelo nulo y el ajustado (550.66), junto con el pvalor asociado al test de razón de verosimilitud ( $p=0.0000$ ).

```
. logit enf sexo edad duracion colesterol
```

```
Iteration 0: log likelihood = -1447.6438
```

```
Iteration 1: log likelihood = -1177.5354
```

```
Iteration 2: log likelihood = -1172.3253
```

```
Iteration 3: log likelihood = -1172.3131
```

```
Iteration 4: log likelihood = -1172.3131
```

```
Logistic regression
```

```
Number of obs = 2258
```

```
LR chi2(4) = 550.66
```

```
Prob > chi2 = 0.0000
```

```
Log likelihood = -1172.3131
```

```
Pseudo R2 = 0.1902
```

```
-----
```

enf		Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
-----	--	-------	-----------	---	------	----------------------



sexo		-2.101717	.1135597	-18.51	0.000	-2.32429 -1.879144
edad		.0730318	.0061459	11.88	0.000	.0609861 .0850775
duracion		-.0016954	.0010139	-1.67	0.094	-.0036826 .0002917
colesterol		.0091233	.0010793	8.45	0.000	.0070079 .0112387
_cons		-4.286432	.3868011	-11.08	0.000	-5.044548 -3.528316

Podemos obtener también el valor del estadístico  $C_g$  de Hosmer y Lemeshow y la significación del test, que en este caso es del orden de 0.5097, lo que sería considerado como un buen ajuste.

```
. estat gof, table group(10)
```

Logistic model for enf, goodness-of-fit test

(Table collapsed on quantiles of estimated probabilities)

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total	
1	0.2863	52	43.7	174	182.3	226	
2	0.4354	75	82.1	151	143.9	226	
3	0.5705	111	113.9	115	112.1	226	
4	0.6674	133	140.7	93	85.3	226	
5	0.7323	157	157.9	68	67.1	225	
6	0.7830	167	171.4	59	54.6	226	
7	0.8289	190	182.3	36	43.7	226	
8	0.8615	193	191.0	33	35.0	226	
9	0.8980	203	198.9	23	27.1	226	

```

|    10 | 0.9796 |    209 | 208.0 |    16 | 17.0 |    225 |
+-----+
      number of observations =      2258
            number of groups =         10
Hosmer-Lemeshow chi2(8) =         7.25
              Prob > chi2 =         0.5097

```

Por último, planteando el modelo dentro de la clase de modelos lineales generalizados, obtendríamos el valor del estadístico Chi Cuadrado de Pearson ( $Pearson = 2275,61$ ) y el de la devianza ( $Deviance = 2344,63$ ).

```
. binreg enf sexo edad duracion colesterol, or
```

```

Iteration 1:  deviance =  2348.799
Iteration 2:  deviance =  2344.634
Iteration 3:  deviance =  2344.626
Iteration 4:  deviance =  2344.626

```

Generalized linear models	No. of obs	=	2258
Optimization : MQL Fisher scoring	Residual df	=	2253
(IRLS EIM)	Scale parameter	=	1
Deviance = 2344.626128	(1/df) Deviance	=	1.040668
Pearson = 2275.614687	(1/df) Pearson	=	1.010038
Variance function: $V(u) = u*(1-u)$	[Bernoulli]		
Link function : $g(u) = \ln(u/(1-u))$	[Logit]		
	BIC	=	-15053.57

```

|                                     EIM

```

enf		Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----							
	sexo	.1222463	.0138822	-18.51	0.000	.097853	.1527206
	edad	1.075765	.0066115	11.88	0.000	1.062884	1.088801
	duracion	.998306	.0010121	-1.67	0.094	.9963242	1.000292
	colesterol	1.009165	.0010892	8.45	0.000	1.007032	1.011302

## 5.4. Conclusiones

La disponibilidad de los seis test estadísticos estudiados se pone de manifiesto en la siguiente tabla, siendo el más completo el lenguaje R.

Estadístico	Software		
	R	SPSS	Stata
$D$	✓	✓	✓
$\chi^2$	✓		✓
$C_g$	✓	✓	✓
$H_g$	✓		
$\hat{T}_{lc}$	✓		
$\hat{S}_{ST}$	✓		

En los modelos planteados se alcanzan resultados similares en los test disponibles en los 3 programas, existiendo alguna diferencia en el estadístico  $C_g$ , tal y cómo se comprueba en la tabla siguiente, dónde se muestran, tanto para el modelo de regresión logística simple construido como para el múltiple, los valores del estadístico, grados de libertad y pvalor para cada uno de los software.

Mod. Reg. Simple				Mod. Reg. Múltiple		
	$C_g$	g.l.	pvalor	$C_g$	g.l.	pvalor
R	9.03	8	0.3399	7.07	8	0.5294
SPSS	8.24	8	0.4100	7.72	8	0.461
Stata	9.00	8	0.3419	7.25	8	0.5097

En cuanto a la comparación de los resultados entre los distintos métodos, el test de razón de verosimilitud y el test basado en el estadístico Chi Cuadrado dan por incorrecto el ajuste de los valores ajustados a los observados, aunque la conclusión no sería válida dado que la aproximación asintótica de los test no es correcta. Entre los otros cuatro test, en el primero de los modelos, si consideramos un nivel de significación del 1 % todos los métodos aceptarían el buen ajuste, sin embargo, si optamos por trabajar a un nivel del 5 % el test basado en  $H_g$  proporciona resultados en dirección contrario. En el modelo de regresión múltiple, trabajando a cualquiera de los niveles de significación comentados, el test  $H_g$  y el test de Stukel darían por inválido el ajuste, mientras que el test de le Cessie y Van Houwelingen y el test de Hosmer Lemeshow  $C_g$  lo consideran aceptable.

A través de estudios de simulación, como los llevados a cabo por Hallet (Hallet [1999]) o Hosmer-Lemeshow (Hosmer et al. [1997]) se ponen de manifiesto el buen o mal funcionamiento de los distintos métodos de bondad de ajuste en distintas situaciones. Como en este trabajo se tomó como referencia el estudio llevado a cabo por el primero de los dos autores anteriores, comentaremos a continuación las conclusiones que obtuvo. En lo que respecta al estadístico basado en la devianza, Hallet comprobó como no rechazaba la hipótesis nula de buen ajuste todas las veces que debería, debido a su pe-

queño error de tipo I, que es considerado como una buena característica en un test de bondad de ajuste. También constata, que al decrecer el número de patrones de las covariables, dicho estadístico puede pasar de rechazar la hipótesis nula a no rechazarla. En situaciones tales que el número de observaciones para cada combinación de valores de las variables explicativas sea un número relativamente pequeño, la aproximación Chi Cuadrado para el estadístico de Pearson y el de la Devianza resulta inadecuada, porque hay pocas réplicas y las probabilidades estimadas por el modelo son cercanas a 0 o a 1. Así, para la correcta aplicación de estos dos estadísticos, deben existir un número suficiente de réplicas. El estadístico basado en la devianza sí que resulta adecuado cuando todas las variables predictoras son categóricas o cuando el experimento está controlado (Simonoff [1998]). Por contra, cuando las variables son continuas, el estadístico  $\chi^2$  pudiera ser de utilidad de cara al cálculo de la media y la varianza condicional.

Hallet considera en su trabajo como mejor método para medir la bondad de ajuste al test score de Brown, dada su alta potencia sobre todo al utilizar la distribución de Cauchy o de valores extremos, especialmente con tamaños de muestra pequeños, y funcionando adecuadamente tanto en modelos logísticos binarios como en binomiales.

Respecto al estadístico  $C_g$ , el autor concluye en su trabajo como que a pesar de estar pensado para un escenario binario, funciona razonablemente bien en el caso binomial, aunque no es del todo fiable en tanto que depende de los grupos de riesgo que se consideren (Hosmer et al. [1997]). Asimismo comprueba como el estadístico  $H_g$  no rechazaba la hipótesis nula tantas veces como debiera, ya que la elección sistemática de 10 grupos puede no ser la adecuada en muchas situaciones, para lo que recomienda la revisión minuciosa de los datos antes de formar y decidir el número de grupos.

En cuanto al estadístico de le Cessie y Van Houwelingen, Hallet destaca su sensibilidad ante tamaños de muestra pequeños, aunque con tamaño de muestra 500 obtiene buenos resultados. La ventaja en el uso de este estadístico radica en que se puede utilizar tanto en modelos binomiales como binarios.







# Bibliografía

Bmdp. URL <http://www.statistical-solutions-software.com/bmdp-statistical-software>.

Gnu. URL <http://gnu.org/copyleft/gpl.html>.

Ibm. URL <http://publib.boulder.ibm.com/infocenter/spssstat/v20r0m0/index.jsp>.

Spss. URL <http://www-01.ibm.com/software/analytics/spss/>.

Stata. URL <http://www.stata.com/>.

A. Agresti. *Categorical Data Analysis*. Wiley, 2002.

L. Silva Aycaguer and I. Barroso Utra. *Regresión Logística*. La Muralla.

A. Azzaline and W. Härdle. On the use of nonparametric regression for model checking. *Biometrika*, 76:1–12, 1989.

E. Bausela. Spss: un instrumento de análisis de datos cuantitativos. *Revista de Informática Educativa y Medios Audiovisuales*, 2:62–69, 2005.

C.C. Brown. On a goodness of fit test for the logistic model based on score statistics. *Communications in Statistics Theory and Methods*, 11:1097–1105, 1982.

- D. Collett. *Modelling binary data*. Chapman and Hall. London., 1991.
- G. Collomb. Estimation non paramétrique de la régression: Revue bibliographique. *International Statistical Review*, 49:75–93, 1981.
- J.B. Copas. Plotting p against x. *Applied Statistics*, 32:25–31, 1983.
- Michael J. Crawley. *Statistics: An Introduction using R*. Springer, 2005.
- Michael J. Crawley. *The R Book*. Wiley.Springer, 2007.
- E.B. Fowlkes. Some diagnostics for binary logistic regression via smoothing. *Biometrika*, 74:503–515, 1987.
- Th. Gasser and H.G. Müller. *Kernel estimations of regression functions. Smoothing Techniques for Curve Estimation*. Berlin Springer-Berlag, 1979.
- T. Gordon, W.B. Kannel, and M. Halpebn. Prediction of coronary disease. *Journal of Chronic Diseases*, 32:427–440, 1979.
- D.C Hallet. *Goodness of fit tests in logistic regression*. PhD thesis, University of Toronto, 1999.
- D.W. Hosmer and S. Lemeshow. Goodness of fit tests for the multiple logisac regression model. *Communications in Statistics*, 10:1043–1069, 1980.
- D.W. Hosmer and S. Lemeshow. *Applied Logistic regression*. Wiley, 1989.
- D.W. Hosmer, T. Hosmer, S. le Cessie, and S. Lemeshow. A comparison of goodness of it tests for the logistic regression model. *Statistics in Medicine*, 16:965–980, 1997.
- D.G. Kleinbaum. *Logistic Regression. A Self-Learning Text*. Springer, 1994.

- M. Kohl. *MKmisc: Miscellaneous functions from M. Kohl. R package version 0.94*.
- J.M. Landwehr, D. Pregiboon, and A.C. Shoemaker. Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association*, 79:61–83, 1984.
- S. le Cessie S. and J.C. Van Houwelingen. A goodness of fit test for binay regression models, based on smoothing methods. *Biometrics*, 47:1267–1282, 1991.
- S. le Cessie S. and J.C. Van Houwelingen. Testing the fit of a regression model via score tests in random effects models. *Biometrics*, 51:600–614, 1995.
- S. Lemeshow and D.W. Hosmer. A review of goodness of fit statistics for use in the development of logistic regression models. *American Journal of Epidemiology*, 115:92–106, 1982.
- J.S. Long and J. Freese. *Regression Models for Categorical Dependent Variables Using Stata*. StataPress, 2006.
- P. McCullagh and J. Nelder. *Generalized Linear Models*. Chapman and Hall, 1989.
- E.A. Nadaraya. On estimation regression. *Theory of Probability and Its Applications*, 9:141–142, 1964.
- J.A. Nelder and R.W.M. Wedderburn. General linear models. *Journal of the Royal Statistical Society Series A*, 135, 1972.
- D.A. Power and Y. Xie. *Statistical Methods for Categorical Data Analysis*. Academia Press, 2000.

- R.L. Prentice. A generalization of the probit and logit methods for dose response curves. *Biometrics*, 32:761–768, 1976.
- B. Presnell. URL <http://www.stat.ufl.edu/presnell/>.
- T.P. Ryan. *Modern regression methods*. Wiley, 1997.
- T.J. Santner and D.E. Duffy. A note on a. albert and j. a. anderson’s conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 73:755–758, 1986.
- S. Selvin. *Statistical Analysis of Epidemiological Data*. Oxford University Press, 1996.
- J.S. Simonoff. Logistic regression, categorical predictors, and goodness-of-fit: It depends on who you ask. *The American Statistician*, 52:10–14, 1998.
- T.A. Stukel. Generalized logistic models. *Journal of the American Statistical Association*, 83:426–431, 1988.
- L. Thompson. *S-PLUS (and R) Manual to Accompany Agresti’s (2002) Categorical Data Analysis*. 2007.
- A.A. Tsiatis. A note on a goodness-of-fit test for the logistic regression model. *Journal Biometrika*, 67:250–251, 1980.
- W. Venables and B. Ripley. *Modern applied statistics with S*. Springer, 2003.
- G.R. Warnes. *gdata: Various R programming tools for data manipulation*. 2012.
- G.S. Watson. Smooth regression analysis. *Sankhya Series A*, 26:359–372, 1964.