concordance=TRUE

Departamento de Estadística e I.O.

Máster en Estadística Aplicada

UNIVERSIDAD DE GRANADA

MODELOS DE RESPUESTA DISCRETA APLICACIONES BIOSANITARIAS

Tema 3 de prácticas
Ajuste de regresión logística con variables explicativas cualitativas con R

Profesores

Ana María Aguilera del Pino

Manuel Escabias Machuca

Título original: Modelos de Respuesta Discreta. Aplicaciones Biosanitarias. Tema 3 de prácticas: Ajuste de regresión logística con variables explicativas cualitativas con R

Contents

Chapter 1

Ajuste de regresión logística con variables explicativas cualitativas con R

1.1 Introducción

Este capítulo tiene por objetivo mostrar el ajuste de modelos de regresión logística con variables explicativas cualitativas. Antes de entrar en las explicaciones de este tipo de ajustes con R recordemos brevemente algunos aspectos teóricos a tener en cuenta.

Hay que decir que el tratamiento de variables explicativas culitativas es común para todos los modelos de regresión: lineal, logística, de Poisson, de respuesta múltiple, etc. Dado que cualquiera de los modelos mencionados son modelos numéricos, el tratamiento de variables cualitativas, en esencia no numéricas, pasa por utilizar como variables explicativas otras variables numéricas auxiliares, de manera que la información que hay en estas variables numéricas sea la misma que contiene la variable cualitativa de interés. Estas variables numéricas se denominan variables de diseño o variables *dummy*.

Por ejemplo, supongamos que nuestra variable cualitativa se denomina *presión arterial* y que puede tomar los valores *baja, normal, alta*. Tal y como se indicó en la teoría, para la creación de las variables de diseño hay que tener en cuenta:

- Como la variable cualitativa tiene tres posibles valores, para representar toda la información de dicha variable se necesitan dos variables de diseño (número de categorías menos una).

- Una de las categorías se considera como categoría de referencia.

- Cada variable de diseño tomará un valor para cada observación de la variable cualitativa.

Dependiendo del modo de codificación los valores que toman las variables de diseño para cada valor de la variable cualitativa serán unos u otros.

Siguiendo con nuestro ejemplo, supongamos que denotamos por X a la variable cualitativa *presión arterial* y por X_1^D y X_2^D a las variables de diseño, entonces si tomamos codificación parcial y como categoría de referencia la categoría *baja*, las variables de diseño tomarían los siguientes valores para cada categoría de la variable cualitativa:

	X: Presión arterial	X_1^D	X_2^D
Cat. Referencia	Baja	0	0
	Normal	1	0
	Alta	0	1

De esta manera, lo que sería el modelo logístico estándar para explicar una variable respuesta Y a partir de la variable explicativa cualitativa X (presión arterial)

$$y = p(x) + \varepsilon,\ p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

se convierte en el modelo logístico para explicar la variable respuesta Y a partir de las variables de diseño

$$y = p(x_1^D, x_2^D) + \varepsilon,\ p(x_1^D, x_2^D) = \frac{e^{\beta_0+\tau_1 x_1^D+\tau_2 x_2^D}}{1 + e^{\beta_0+\tau_1 x_1^D+\tau_2 x_2^D}}$$

Si bien teóricamente éste es el método para incluir variables cualitativas en modelos de regresión, todos los programas hacen esto de manera automática y transparente para el usuario como se verá más adelante. Aun así conocer los aspectos teóricos es muy útil para hacer comprobaciones y posibles modificaciones.

Reflesiónese cómo para los modelos de regresión logística, al tomar como variable respuesta una variable de ceros y unos, realmente lo que se está utilizando es una variable de diseño que representa a una variable cualitativa que toma los valores *éxito, fracaso* con categoría de referencia el *fracaso* y codificación parcial.

Para ilustrar todas las cuestiones del tratamiento de variables cualitativas en regresión logística, se utilizará el mismo conjunto de datos del Tema 2 en el que se han sustituido las variables *Sistólica* y *Diastólica* por una variable cualitativa llamada *Presión* con los valores *óptima, normal, alta* y *descompensada*, y las variables *Peso* y *Altura* por la variable cualitativa *IMC* que toma los valores *normal, sobrepeso* y *obesidad*. Los datos están

disponibles en el fichero chapman_Cuali.csv, que se trata de un fichero de texto plano, configurado en 6 columnas separadas por comas. La primera columna contiene un indicador numérico que identifica cada caso.

1.2 Analísis de regresión logística y variables explicativas cualitativas con R

Antes de entrar en el ajuste de regresión logística con variables explicativas cualitativas se explicará cómo R realiza la lectura y tratamiento de variables cualtitativas, apoyados en el conjunto de datos *Chapman_Cuali.csv*.

Para leer este conjunto de datos, recurrimos como siempre a la sentencia `read.csv()` (si se utilizar RStudio se puede utilizar el ratón)

```
Chapman.Cuali<-read.csv("Chapman_Cuali.csv",header=T,sep=",")
```

De esta manera se genera un *Data.Frame* con tres columnas numéricas (Id, Edad, Colesterol y Coronarios) y dos columnas no numéricas (Presión e IMC). Este hecho se puede comprobar con las sentencias siguientes:

```
is.numeric(Chapman.Cuali$Id)

## [1] TRUE

is.numeric(Chapman.Cuali$Edad)

## [1] TRUE

is.numeric(Chapman.Cuali$Colesterol)

## [1] TRUE

is.factor(Chapman.Cuali$Presion)

## [1] TRUE

is.factor(Chapman.Cuali$IMC)

## [1] TRUE

is.numeric(Chapman.Cuali$Coronarios)

## [1] TRUE
```

Así mismo se puede comprobar que (ID, Edad, Colesterol y Coronarios) no son de tipo factor (Presión e IMC) no son numéricas.

```
is.factor(Chapman.Cuali$ID)

## [1] FALSE

is.factor(Chapman.Cuali$Edad)

## [1] FALSE

is.factor(Chapman.Cuali$Colesterol)

## [1] FALSE

is.numeric(Chapman.Cuali$Presion)

## [1] FALSE

is.numeric(Chapman.Cuali$IMC)

## [1] FALSE

is.factor(Chapman.Cuali$Coronarios)

## [1] FALSE
```

Hay que decir que con la sentencia `read.csv()` cuando el programa se encuentra en el fichero de texto una columna en la que todos los valores son numéricos (o en blanco), se lee como variable numérica, mientras que si se encuentra al menos un valor no numérico, considera toda la variable como no numérica (cualitativa). Además por defecto el programa R trata a las variables no numéricas como de tipo `factor` a no ser que se indique lo contrario. Dado que para el ajuste de modelos es conveniente que las variables no numéricas sean de tipo `factor` no se modificará nada en este aspecto.

El programa R trata a las variables de tipo `factor` como si fueran de tipo entero, en el sentido de que les asigna un orden. Por defecto el orden que se asigna es el alfabético como se puede ver con la sentencia

```
levels(Chapman.Cuali$Presion)

## [1] "Alta"          "Descompensada" "Normal"        "Optima"

levels(Chapman.Cuali$IMC)

## [1] "Normal"    "Obesidad"  "Sobrepeso"
```

Este hecho también se aprecia con las siguentes sentencias

```
contrasts(Chapman.Cuali$Presion)

##               Descompensada Normal Optima
## Alta                      0      0      0
## Descompensada             1      0      0
## Normal                    0      1      0
## Optima                    0      0      1

contrasts(Chapman.Cuali$IMC)

##           Obesidad Sobrepeso
## Normal           0         0
## Obesidad         1         0
## Sobrepeso        0         1
```

con las que además de apreciar el orden se puede ver el tratamiento que tendrían estas variables en su inclusión en un modelo de regresión, esto es, la codificación de las variables de diseño. Se puede ver por ejemplo, que (por defecto) la variable *Presión* tendría como categoría de referencia la categoría *Alta* mientras que la variable *IMC* tiene como categoría de referencia la categoría *Normal*.

Estas características que toma R por defecto se pueden cambiar. Supongamos que a las categorías de la variable presión queremos que la categoría de referencia sea la categoría *Optima* y que para el IMC queremos que sea *Normal*, las sentencias serían

```
Chapman.Cuali$Presion<-relevel(Chapman.Cuali$Presion,ref="Optima")
Chapman.Cuali$IMC<-relevel(Chapman.Cuali$IMC,ref="Normal")
```

En cuyo caso vemos cómo cambia la categoría de referencia y la asignación de las variables de diseño para su uso en modelos de regresión.

```
contrasts(Chapman.Cuali$Presion)

##               Alta Descompensada Normal
## Optima           0             0      0
## Alta             1             0      0
## Descompensada    0             1      0
## Normal           0             0      1

contrasts(Chapman.Cuali$IMC)

##           Obesidad Sobrepeso
## Normal           0         0
## Obesidad         1         0
## Sobrepeso        0         1
```

De ahora en adelante se asumirá estas categorías de referencia para las variables cualitativas de este ejemplo, en lugar del que asigna por defecto R, esto es, asumimos que hemos cambiado la categoría de referencia por defecto con las sentencias `relevel()<-`.

1.3 Ajuste de regresión logística simple para una variable explicativa cualitativa

Supongamos que se quiere modelizar la ocurrencia de enfermedad coronaria de nuestro ejemplo a partir de la variable cualitativa *Presion*.

```
Ajuste.Presion<-glm(Coronarios~Presion,data=Chapman.Cuali,family=binomial)
summary(Ajuste.Presion)

##
## Call:
## glm(formula = Coronarios ~ Presion, family = binomial, data = Chapman.Cual
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8346  -0.5553  -0.4761  -0.2649   2.5951
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -2.1203     0.4320  -4.907 9.23e-07 ***
```

```
## PresionAlta            1.2448     0.6856   1.816   0.0694 .
## PresionDescompensada   0.3285     0.5196   0.632   0.5273
## PresionNormal         -1.2119     1.1056  -1.096   0.2730
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 154.55  on 199  degrees of freedom
## Residual deviance: 147.82  on 196  degrees of freedom
## AIC: 155.82
##
## Number of Fisher Scoring iterations: 6
```

El resultado del ajuste nos muestra los parámetros asociados a cada una de las variables de diseño que automática e internamente ha utilizado el programa. Obsérvese cómo la variable *Presion* tenía cuatro categorías (*Optima,Normal,Alta,Descompensada* en ese orden) y por tanto para el ajuste se utilizan tres variables de diseño llamadas `PresionAlta, PresionDescompensada, PresionNormal` que por otro lado indica que la categoría de referencia es *Optima*. La ecuación de predicción de este modelo será

$$\begin{aligned}\hat{p} &= \frac{exp(\hat{\beta}_0 + \hat{\tau}_1^P X_1^P + \hat{\tau}_2^P X_2^P + \hat{\tau}_3^P X_3^P)}{1 + exp(\hat{\beta}_0 + \hat{\tau}_1^P X_1^P + \hat{\tau}_2^P X_2^P + \hat{\tau}_3^P X_3^P)} = \\ &= \frac{exp(-2.12 + (1.24)X_1^P + (0.33)X_2^P + (-1.21)X_3^P)}{1 + exp(-2.12 + (1.24)X_1^P + (0.33)X_2^P + (-1.21)X_3^P)}\end{aligned}$$

siendo $X_j^P,\ j = 1, 2, 3$ las variables de diseño con codificación parcial y $\hat{\tau}_j^P$ los parámetros estimados asociados.

Una vez ajustado el modelo se pueden estudiar todos los aspectos vistos en el tema 2:

- Predicción e interpretación de parámetros.
- Bondad del ajuste.
- Significación de parámetros.
- Validación.
- Precisión del modelo logístico y clasificación de observaciones.

En este punto nos centraremos en la interpretación de los parámetros puesto que el resto apartados no difieren mucho de lo visto en el Tema 2.

Recuerdese que la exponencial de los parámetros del modelo logístico representa el cambio multiplicativo que ocurre en la ventaja de respuesta 1 (frente a respuesta 0) al cambiar el valor de la variable explicativa. En este caso (codificación parcial), si llamamos $\hat{p}_1, \hat{p}_2,\ \hat{p}_3,\ \hat{p}_4$ a la probabilidad de éxito (padecer problemas coronarios) para cada una de las categorías de presión (óptima, normal, alta y descompensada), se tendría que:

- La exponencial del parámetro asociado a la variable de diseño asociada con la categoría *Presión Alta* es el cociente de ventajas

$$\hat{\theta}_{21} = \frac{\dfrac{\hat{p}_2}{1-\hat{p}_2}}{\dfrac{\hat{p}_1}{1-\hat{p}_1}} = \exp\left(\hat{\tau}_1^P\right) = e^{1.2447948} = 3.4722222.$$

  Esto significa que la ventaja de padecer problemas coronarios (frente a no padecerlos) si tienes la presión alta es 3.4722222 veces la ventaja cuando tienes la presión óptima. Dicho de otro modo, la ventaja se multiplica por 3.4722222 cuando se pasa de Presión óptima a presión alta.

- La exponencial del parámetro asociado a la variable de diseño asociada con la categoría *Presión Descompensada* es el cociente de ventajas

$$\hat{\theta}_{31} = \frac{\dfrac{\hat{p}_3}{1-\hat{p}_3}}{\dfrac{\hat{p}_1}{1-\hat{p}_1}} = \exp\left(\hat{\tau}_2^P\right) = e^{0.3285041} = 1.3888889.$$

  Esto significa que la ventaja de padecer problemas coronarios (frente a no padecerlos) si tienes la presión descompensada es 1.3888889 veces la ventaja cuando tienes la presión óptima. Dicho de otro modo, la ventaja se multiplica por 1.3888889 cuando se pasa de Presión óptima a presión descompensada.

- La exponencial del parámetro asociado a la variable de diseño asociada con la categoría *Presión Normal* es el cociente de ventajas

$$\hat{\theta}_{41} = \frac{\dfrac{\hat{p}_4}{1-\hat{p}_4}}{\dfrac{\hat{p}_1}{1-\hat{p}_1}} = \exp\left(\hat{\tau}_3^P\right) = e^{-1.211941} = 0.297619.$$

Esto significa que la ventaja de padecer problemas coronarios (frente a no padecerlos) si tienes la presión normal es 0.297619 veces la ventaja cuando tienes la presión óptima. Dicho de otro modo, la ventaja se multiplica por 0.297619 cuando se pasa de Presión óptima a presión normal.

- La exponencial del parámetro independiente es la ventaja

$$\frac{\hat{p}_0}{1-\hat{p}_0} = exp\widehat{\beta}_0 = e^{-2.1202635} = 0.12.$$

Esto significa que la ventaja de padecer problemas coronarios (frente a no padecerlos) si tienes la presión óptima es 0.12. Esto es, es 0.12 veces menos probable padecer problemas coronarios que no padecerlos.

Obsérvese que las interpretaciones aportadas lo han sido de los parámetros estimados, independientemente de la significación de tales parámetros. Mirando la significación de los parámetros, sólo el de la presión alta es significativo, y por tanto el resto se deberían interpretar como nulos y su exponencial (odds ratio) como 1, no produciéndose cambio alguno en las ventajas.

A partir de estas interpretaciones se puede obtener cualquier otra que se quiera. Por ejemplo, podríamos estar interesados en conocer por cuánto se multiplica la ventaja de padecer problemas coronarios (frente a no padecerlos) si la presión cambia de Normal a Alta, o de Normal a Descompensada, o de Alta a Descompensada.

Si en lugar de utilizar codificación parcial para las variables de diseño, se utiliza codificación marginal, utilizando también la primera categoría como categoría de referencia, el ajuste que se obtiene es diferente, los parámetros se interpretan de manera diferente, pero las conclusiones sobre el cambio en las ventajas son las mismas.

Para cambiar el sistema de codificación, sin modificar nuestra variable explicativa *presion* creamos una nueva variable llamada *PresionM* que tiene los mismos valores anteriores con las mismas característica de codificación

```
Chapman.Cuali$PresionM<-Chapman.Cuali$Presion
```

Podemos comprobar que tiene el mismo estatus que la variable presion:

```
contrasts(Chapman.Cuali$PresionM)
```

```
##               Alta Descompensada Normal
## Optima           0             0      0
```

```
## Alta             1             0      0
## Descompensada    0             1      0
## Normal           0             0      1
```

Para cambiar el estatus de codificación a marginal y con la última categoría (*Normal*) como referencia utilizamos la siguiente sentencia:

```
contrasts(Chapman.Cuali$PresionM)<-contr.sum(4)
```

tras la que se puede comprobar el cambio en la codificación

```
contrasts(Chapman.Cuali$PresionM)

##               [,1] [,2] [,3]
## Optima           1    0    0
## Alta             0    1    0
## Descompensada    0    0    1
## Normal          -1   -1   -1
```

Si se ajusta el modelo logístico simple, la estimación de parámetros cambia con respecto a la vista previamente:

```
Ajuste.PresionM<-glm(Coronarios~PresionM,data=Chapman.Cuali,family=binomial)
summary(Ajuste.PresionM)

##
## Call:
## glm(formula = Coronarios ~ PresionM, family = binomial, data = Chapman.Cua
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8346  -0.5553  -0.4761  -0.2649   2.5951
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.02992    0.31514  -6.441 1.18e-10 ***
## PresionM1   -0.09034    0.43892  -0.206   0.8369
## PresionM2    1.15446    0.49090   2.352   0.0187 *
## PresionM3    0.23816    0.37548   0.634   0.5259
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 154.55  on 199  degrees of freedom
## Residual deviance: 147.82  on 196  degrees of freedom
## AIC: 155.82
##
## Number of Fisher Scoring iterations: 6
```

La ecuación de predicción de este modelo será

$$\begin{aligned}\hat{p} &= \frac{exp(\hat{\beta}_0 + \hat{\tau}_1^M X_1^M + \hat{\tau}_2^M X_2^M + \hat{\tau}_3^M X_3^M)}{1 + exp(\hat{\beta}_0 + \hat{\tau}_1^M X_1^M + \hat{\tau}_2^M X_2^M + \hat{\tau}_3^M X_3^M)} = \\ &= \frac{exp(-2.03 + (-0.09)X_1^M + (1.15)X_2^M + (0.24)X_3^M)}{1 + exp(-2.03 + (-0.09)X_1^M + (1.15)X_2^M + (0.24)X_3^M)}\end{aligned}$$

siendo $X_j^M,\ j = 1, 2, 3$ las variables de diseño con codificación marginal y $\hat{\tau}_j^M$ los parámetros estimados asociados.

Cuando se usa el método marginal la exponencial de cada parámetro estimado ya no es el cociente de ventajas tal y como se ha visto en el método parcial. Ahora si recordamos que $\hat{p}_1, \hat{p}_2,\ \hat{p}_3,\ \hat{p}_4$ son las probabilidades de éxito (padecer problemas coronarios) para cada una de las categorías de presión (óptima, alta, descompensada y normal), se tendría que:

- El cociente de ventajas

$$\hat{\theta}_{14} = \frac{\frac{\hat{p}_1}{1-\hat{p}_1}}{\frac{\hat{p}_4}{1-\hat{p}_4}} = \exp\left(2 * \hat{\tau}_1 + \hat{\tau}_2 + \hat{\tau}_3\right) = e^{2*(-0.09)+(1.15)+(0.24)} = 3.36.$$

  Esto significa que la ventaja de padecer problemas coronarios (frente a no padecerlos) si tienes la presión óptima es 3.36 veces la ventaja cuando tienes la presión normal. Dicho de otro modo, la ventaja se multiplica por 3.36 cuando se pasa de Presión normal a óptima.

- El cociente de ventajas

$$\hat{\theta}_{24} = \frac{\frac{\hat{p}_2}{1-\hat{p}_2}}{\frac{\hat{p}_4}{1-\hat{p}_4}} = \exp\left(\hat{\tau}_1^M + 2 * \hat{\tau}_2^M + \hat{\tau}_3^M\right) = e^{(-0.09)+2*(1.15)+(0.24)} = 11.67.$$

Esto significa que la ventaja de padecer problemas coronarios (frente a no padecerlos) si tienes la presión alta es 11.67 veces la ventaja cuando tienes la presión normal. Dicho de otro modo, la ventaja se multiplica por 11.67 cuando se pasa de Presión de normal a alta.

- El cociente de ventajas

$$\hat{\theta}_{34} = \frac{\frac{\hat{p}_3}{1-\hat{p}_3}}{\frac{\hat{p}_4}{1-\hat{p}_4}} = \exp\left(\hat{\tau}_1^M + \hat{\tau}_2^M + 2 * \hat{\tau}_3^M\right) = e^{-0.09+(1.15)+2*(0.24)} = 4.67.$$

  Esto significa que la ventaja de padecer problemas coronarios (frente a no padecerlos) si tienes la presión descompensada es 4.67 veces la ventaja cuando tienes la presión normal. Dicho de otro modo, la ventaja se multiplica por 4.67 cuando se pasa de Presión normal a descompensada.

- La ventaja

$$\frac{\hat{p}_4}{1-\hat{p}_4} = \exp\left(\hat{\beta}_0 - \hat{\tau}_1^M - \hat{\tau}_2^M - *\hat{\tau}_3^M\right) = e^{-2.03--0.09-(1.15)-(0.24)} = 0.04.$$

  Esto significa que la ventaja de padecer problemas coronarios (frente a no padecerlos) si tienes la presión descompensada es 0.04.

Como se indicó previamente y se puede comprobar, es posible obtener unos cocientes de ventajas a partir de otros, por ejemplo:

$$\hat{\theta}_{21} = \frac{\hat{\theta}_{24}}{\hat{\theta}_{14}} = \frac{11.6666667}{3.36} = 3.4722222$$

Se deja al lector la comprobación del resto.

Recuerdese que las variables *Presion* y *PresionM* son exáctamente las mismas, esto es, el primer caso de la matriz de datos es una persona que no tenía problemas coronarios y tenía la presión alta (tanto *Presion* como *PresionM*), y el segundo es de una persona que no tenía problemas coronarios y la presión descompensada (tanto *Presion* como *PresionM*). Lo que ha cambiado en todos los pasos explicados previamente es la forma de definir las variables de diseño para los ajustes.

Como ejercicio, obsérvese cómo son los modelos ajustados (parámetros estimados) si no se cambia el orden de las categorías ni la categoría de referencia con respecto a los que por defecto utiliza R.

1.3.1 Otros resultados del ajuste

Predicción

En la predicción de la respuesta por el modelo simple con variables explicativas cualitativas hay que tener presente algunas cuestiones que se comentan a continuación.

Aunque en el conjunto de datos `Chapman.Cuali` hay 200 casos, la variable explicativa presión tiene cuatro categorías, cada una de las cuales tiene un número de éxitos y de fracasos:

```
table(Chapman.Cuali$Presion,Chapman.Cuali$Coronarios)

##
##                   0  1
##   Optima         50  6
##   Alta           12  5
##   Descompensada 84 14
##   Normal         28  1
```

Los valores predichos por el modelo dependen únicamente de las observaciones de la variable explicativa y no de la codificación de las variables de diseño. Para el ajuste realizado con la codificación parcial los 10 primeros valores ajustados son

```
fitted.values(Ajuste.Presion)[1:10]

##          1          2          3          4          5          6
## 0.03448276 0.10714286 0.14285714 0.14285714 0.29411765 0.29411765
##          7          8          9         10
## 0.14285714 0.10714286 0.03448276 0.03448276
```

y para el ajuste realizado con la codificación marginal

```
fitted.values(Ajuste.PresionM)[1:10]

##          1          2          3          4          5          6
## 0.03448276 0.10714286 0.14285714 0.14285714 0.29411765 0.29411765
##          7          8          9         10
## 0.14285714 0.10714286 0.03448276 0.03448276
```

Además como sólo hay cuatro observaciones diferentes de la variable explicativa, el número de valores predichos diferentes, será igual al número

de categorías, cuatro, cada uno de los cuales se repetirá tantas veces como su respectiva categoría.

```
table(fitted.values(Ajuste.Presion),Chapman.Cuali$Coronarios)
```

```
##
##                        0  1
##   0.0344827586207024 28  1
##   0.107142857142857  50  6
##   0.142857142857142  84 14
##   0.294117647058823  12  5
```

La función `predict()` funciona de manera análoga a la función `fitted.values()` con lo que todo lo explicado anteriormente es válido para esta función también, incluídos los errores estándar de estimación de las probabilidades predichas.

Bondad del ajuste

En el caso de regresión logística simple con variables explicativas cualitativas el ajuste siempre es perfecto pues es un modelo con tantos parámetros como observaciones, un modelo saturado, y por lo que no tiene sentido estudiar la bondad del ajuste a través del test de Hosmer y Lemeshow, del test chi-cuadrado o del estadístico de wilks. De hecho se puede comprobar que las probabilidades estimadas coinciden con las frecuencias relativas (proporción de éxitos) en cada categoría de la variable explicativa.

Significación de parámetros

El análisis de la significación de parámetros se lleva a cabo de igual modo que en cualquier otro ajuste de regresión logística. Hay que tener en cuenta que la significación de parámetros tiene que ver con los parámetros del modelo y no con las variables, por lo que depende de la codificación hecha de las variables del diseño para el ajuste.

```
summary(Ajuste.Presion)
```

```
##
## Call:
## glm(formula = Coronarios ~ Presion, family = binomial, data = Chapman.Cual
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -0.8346  -0.5553  -0.4761  -0.2649   2.5951
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -2.1203     0.4320  -4.907 9.23e-07 ***
## PresionAlta           1.2448     0.6856   1.816   0.0694 .
## PresionDescompensada  0.3285     0.5196   0.632   0.5273
## PresionNormal        -1.2119     1.1056  -1.096   0.2730
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 154.55  on 199  degrees of freedom
## Residual deviance: 147.82  on 196  degrees of freedom
## AIC: 155.82
##
## Number of Fisher Scoring iterations: 6
```

```
summary(Ajuste.PresionM)
```

```
##
## Call:
## glm(formula = Coronarios ~ PresionM, family = binomial, data = Chapman.Cuali)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8346  -0.5553  -0.4761  -0.2649   2.5951
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.02992    0.31514  -6.441 1.18e-10 ***
## PresionM1   -0.09034    0.43892  -0.206   0.8369
## PresionM2    1.15446    0.49090   2.352   0.0187 *
## PresionM3    0.23816    0.37548   0.634   0.5259
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 154.55  on 199  degrees of freedom
```

```
## Residual deviance: 147.82  on 196  degrees of freedom
## AIC: 155.82
##
## Number of Fisher Scoring iterations: 6
```

Como puede observarse puede ocurrir que algunas variables de diseño tengan parámetros significativos y otras no significativos. Por esta razón no se puede utilizar en este caso la significación de parámetros como estrategia para seleccionar variables, ya que no se puede eliminar sólo una parte de una variable explicativa cualitativa si ocurriese que sólo algunos parámetros fuesen no significativos.

De los intervalos de confianza de los parámetros se puede hacer la misma reflexión que de la significación de los parámetros.

```
confint.default(Ajuste.Presion)
```

```
##                              2.5 %     97.5 %
## (Intercept)          -2.96706476 -1.2734623
## PresionAlta          -0.09888881  2.5884784
## PresionDescompensada -0.68992309  1.3469312
## PresionNormal        -3.37890260  0.9550206
```

```
confint.default(Ajuste.PresionM)
```

```
##                  2.5 %     97.5 %
## (Intercept) -2.6475940 -1.4122542
## PresionM1   -0.9506024  0.7699234
## PresionM2    0.1923104  2.1166002
## PresionM3   -0.4977545  0.9740837
```

Como la exponencial de los parámetros tienen interpretación en términos de cocientes de ventajas sólo para la codificación parcial, sólo tendrá sentido calcular los intervalos de confianza para los cocientes de ventaja mediate la exponencial de los intervalos de los parámetros en esta codificación.

```
exp(confint.default(Ajuste.Presion))
```

```
##                              2.5 %    97.5 %
## (Intercept)          0.05145412  0.279861
## PresionAlta          0.90584342 13.309505
## PresionDescompensada 0.50161464  3.845606
## PresionNormal        0.03408484  2.598724
```

Si se quisiese estimar intervalos de confianza para otro tipo de cociente de ventajas, habría que seleccionar codificación parcial con la categoría de referencia adecuada y ajustar el modelo logístico adecuado, para posteriormente obtener los intervalos de confianza.

Test condicionales de razón de verosimilitudes. En el caso de variables explicativas cuantitativas se pueden utilizar los test condicionales de razón de verosimilitudes para estudiar la significación de parámetros puesto que cada variable tiene asociado un único parámetro, y es equivalente analizar la significación de un parámetro con analizar la significación de una variable. En el caso de variables explicativas cualitativas esta identificación no es posible pues cada variable tiene asociados varios parámetros, dependiendo del número de variables de diseño. Entonces la significación de parámetros se estudia con el test de Wald, mientras que la significación de variables se estudia con los test condicionales de razón de verosimilitudes.

```
#MODELO GENERAL
Ajuste.Presion

##
## Call:  glm(formula = Coronarios ~ Presion, family = binomial, data = Chapman.Cual
##
## Coefficients:
##          (Intercept)           PresionAlta  PresionDescompensada
##              -2.1203                1.2448                0.3285
##        PresionNormal
##              -1.2119
##
## Degrees of Freedom: 199 Total (i.e. Null);  196 Residual
## Null Deviance:     154.6
## Residual Deviance: 147.8  AIC: 155.8

#MODELO PARTICULAR
Ajuste.0<-glm(Coronarios~1,data=Chapman.Cuali,family=binomial)
#COMPARACION
anova(Ajuste.0,Ajuste.Presion)

## Analysis of Deviance Table
##
## Model 1: Coronarios ~ 1
## Model 2: Coronarios ~ Presion
##   Resid. Df Resid. Dev Df Deviance
```

```
## 1       199     154.56
## 2       196     147.82  3   6.7391

pchisq(anova(Ajuste.0,Ajuste.Presion)$Deviance[2],
       anova(Ajuste.0,Ajuste.Presion)$Df[2],lower.tail=F)

## [1] 0.08069562
```

Si en lugar de utilizar el ajuste de la codificación parcial, se utiliza el de la marginal, el resultado es el mismo puesto que el valor de la *deviance* coincide en ambos ajustes. Veamos:

```
anova(Ajuste.0,Ajuste.PresionM)

## Analysis of Deviance Table
##
## Model 1: Coronarios ~ 1
## Model 2: Coronarios ~ PresionM
##   Resid. Df Resid. Dev Df Deviance
## 1       199     154.56
## 2       196     147.82  3   6.7391

pchisq(anova(Ajuste.0,Ajuste.PresionM)$Deviance[2],
       anova(Ajuste.0,Ajuste.PresionM)$Df[2],lower.tail=F)

## [1] 0.08069562
```

Validación

La definición de residuos hace que con los residuos ocurra algo similar a lo que ocurre con los valores predichos por el modelo. En este caso en lugar de haber tantos residuos diferentes como categorías de la variable explicativa, hay el doble (cuatro para los éxitos y cuatro para los fracasos). Además coinciden los residuos en los dos tipos de modelos: codificación parcial y marginal.

```
residuals(Ajuste.Presion,type= "pearson")[1:10]

##          1          2          3          4          5          6
## -0.1889822 -0.3464102 -0.4082483 -0.4082483  1.5491933 -0.6454972
```

```
##          7          8          9         10
## -0.4082483 -0.3464102 -0.1889822 -0.1889822

table(residuals(Ajuste.Presion,type= "pearson"))

##
## -0.645497224367903 -0.408248290463862 -0.346410161513775
##                 12                 84                 50
##  -0.18898223650465   1.54919333848297   2.44948974278318
##                 28                  5                 14
##   2.88675134594814   5.29150262212817
##                  6                  1

residuals(Ajuste.Presion,type= "deviance")[1:10]

##          1          2          3          4          5          6
## -0.2649201 -0.4760855 -0.5552489 -0.5552489  1.5644650 -0.8346337
##          7          8          9         10
## -0.5552489 -0.4760855 -0.2649201 -0.2649201

table(residuals(Ajuste.Presion,type= "deviance"))

##
## -0.834633685239477 -0.555248916842271 -0.476085465661372
##                 12                 84                 50
##  -0.26492006270301   1.5644650405951   1.97276970224875
##                 28                  5                 14
##   2.11357148992273   2.59510918074215
##                  6                  1
```

Para detectar valores mál ajustados, los residuos estandarizados

```
rstandard(Ajuste.Presion,type= "pearson")[1:10]

##          1          2          3          4          5          6
## -0.1923273 -0.3495452 -0.4103473 -0.4103473  1.5968719 -0.6653633
##          7          8          9         10
## -0.4103473 -0.3495452 -0.1923273 -0.1923273

table(rstandard(Ajuste.Presion,type= "pearson"))

##
```

```
## -0.665363309277971 -0.410347267232383  -0.34954515900212
##                 12                  84                  50
## -0.192327314540555   1.59687194226713   2.46208360339431
##                 28                   5                  14
##   2.91287632501768   5.38516480713348
##                  6                   1

rstandard(Ajuste.Presion,type= "deviance")[1:10]

##          1          2          3          4          5          6
## -0.2696093 -0.4803940 -0.5581037 -0.5581037  1.6126137 -0.8603207
##          7          8          9         10
## -0.5581037 -0.4803940 -0.2696093 -0.2696093

table(rstandard(Ajuste.Presion,type= "deviance"))

##
## -0.860320710735222 -0.558103685874799 -0.480394019234293
##                 12                  84                  50
## -0.269609277411275   1.61261365248996   1.98291254392473
##                 28                   5                  14
##   2.13269922359946   2.64104388276407
##                  6                   1
```

y para valores influyentes, los valores hat y distancias de Cook

```
hatvalues(Ajuste.Presion)[1:10]

##          1          2          3          4          5          6
## 0.03448276 0.01785714 0.01020408 0.01020408 0.05882353 0.05882353
##          7          8          9         10
## 0.01020408 0.01785714 0.03448276 0.03448276

table(hatvalues(Ajuste.Presion))

##
## 0.0102040816326531 0.0178571428571427 0.0178571428571429
##                 98                  1                  55
## 0.0344827586206897 0.0344827586206901 0.0588235294117647
##                 28                  1                  17

cooks.distance(Ajuste.Presion)[1:10]
```

```
##            1            2            3            4            5
## 0.0003302660 0.0005553719 0.0004339816 0.0004339816 0.0398437500
##            6            7            8            9           10
## 0.0069173177 0.0004339816 0.0005553719 0.0003302660 0.0003302660
```

```
table(hatvalues(Ajuste.Presion))
```

```
##
## 0.0102040816326531 0.0178571428571427 0.0178571428571429
##                 98                  1                 55
## 0.0344827586206897 0.0344827586206901 0.0588235294117647
##                 28                  1                 17
```

Tabla de clasificación

La tasa de clasificaciones correctas depende de las probabilidades predichas por lo que todo lo dicho en la explicación de aquellas y no cambia nada en relación a los modelos con variables explicativas cuantitativas

En nuestro ejemplo la tabla de clasificación ha sido

	Clasificación	
Observación	Fracaso	éxito
Fracaso	VN=78	FP=96
éxito	FN=7	VP=19

Y las medidas:

- Tasa de clasificaciones correctas: CCR =48.5%
- Sensibilidad: 73.08%
- Especificidad: 44.83%

Curva ROC

La Curva ROC también depende de las probabilidades predichas por lo que todo lo dicho en la explicación de aquellas y no cambia nada en relación a los modelos con variables explicativas cuantitativas

```
library(pROC)
CurvaROC<-roc(Chapman.Cuali$Coronarios,fitted.values(Ajuste.Presion))
CurvaROC
```

```
##
## Call:
## roc.default(response = Chapman.Cuali$Coronarios, predictor = fitted.values
##
## Data: fitted.values(Ajuste.Presion) in 174 controls (Chapman.Cuali$Coronar
## Area under the curve: 0.6304
```

```
CurvaROC$thresholds
```

```
## [1]       -Inf 0.07081281 0.12500000 0.21848739        Inf
```

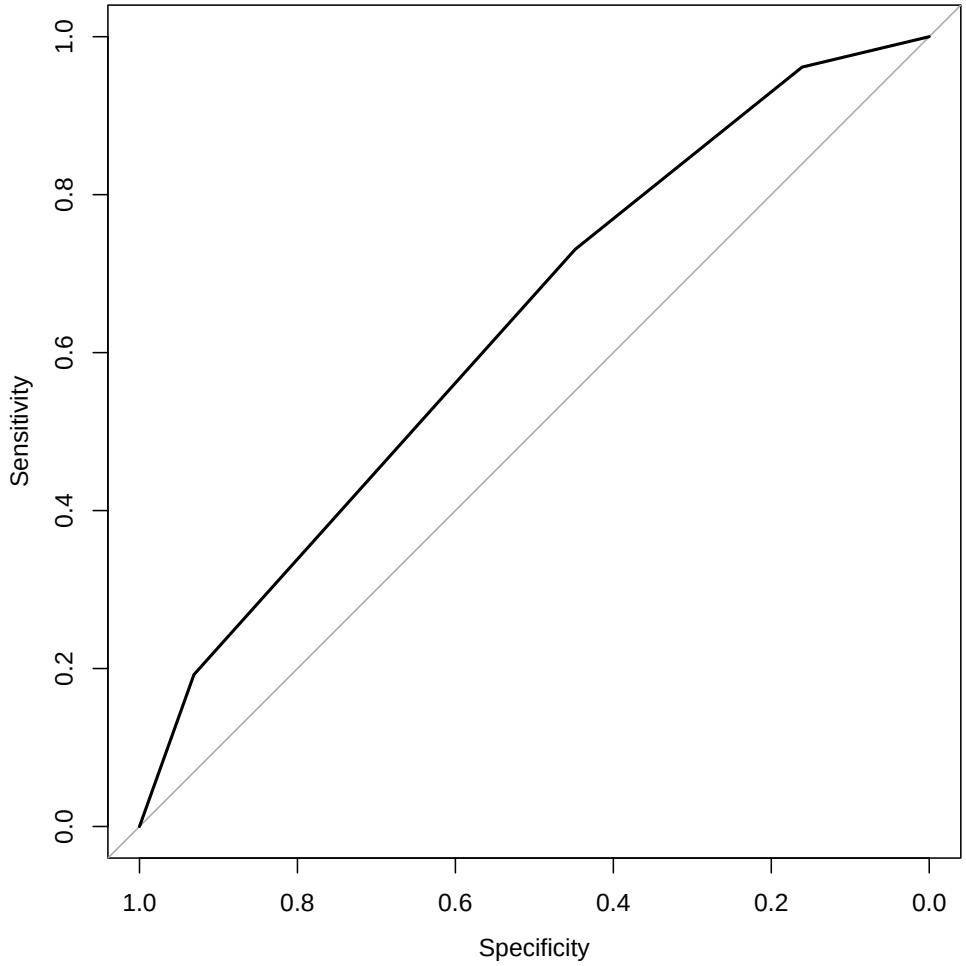
```
CurvaROC$sensitivities
```

```
## [1] 1.0000000 0.9615385 0.7307692 0.1923077 0.0000000
```

```
CurvaROC$specificities
```

```
## [1] 0.0000000 0.1609195 0.4482759 0.9310345 1.0000000
```

```
plot(CurvaROC)
```



Debido a las pocas probabilidades predichas diferentes, la curva ROC no proporciona un método diagnóstico efectivo para el modelo logístico.

1.4 Ajuste del modelo múltiple y selección stepwise

Para ilustrar el ajuste de modelo múltiple que incluya tanto variables cualitativas como cuantitativas utilizaremos el modelo que resulta de la selección stepwise de variables, de este se explican ambos procedimientos a la vez.

Continuaremos en esta ilustración con el ejemplo visto al principio de este documento, cuya información se encuentra en el conjunto de datos *Chapman_Cuali.csv*. Se asume aquí, que se han leído el conjunto de datos

en R con la sentencia `read.csv()` (el ratón en RStudio). Recuérdese que este conjunto contiene tres variables numéricas (Id, Edad, Colesterol y Coronarios) y dos no numéricas (Presión e IMC). Se asume en este punto que para las variables no numéricas se ha cambiado la categoría de referencia que el programa hace por defecto:

```
contrasts(Chapman.Cuali$Presion)

##               Alta Descompensada Normal
## Optima           0             0      0
## Alta             1             0      0
## Descompensada    0             1      0
## Normal           0             0      1

contrasts(Chapman.Cuali$IMC)

##           Obesidad Sobrepeso
## Normal           0         0
## Obesidad         1         0
## Sobrepeso        0         1
```

El primer paso es la selección stepwise de variables.

```
Ajuste.0<-glm(Coronarios~1,data=Chapman.Cuali,family=binomial)
Ajuste.All<-glm(Coronarios~Edad+Colesterol+Presion+IMC,
                data=Chapman.Cuali,family=binomial)
Ajuste.Step<-step(Ajuste.0,scope=list(lower=Coronarios~1,
upper=Coronarios~Edad+Colesterol+Presion+IMC),
direction = "both")

## Start:  AIC=156.55
## Coronarios ~ 1
##
##              Df Deviance    AIC
## + Edad        1   142.74 146.74
## + Colesterol  1   146.94 150.94
## + IMC         2   146.50 152.50
## + Presion     3   147.82 155.82
## <none>            154.56 156.56
##
## Step:  AIC=146.74
```

```
## Coronarios ~ Edad
##
##              Df Deviance    AIC
## + IMC         2   137.45 145.45
## + Colesterol  1   139.93 145.93
## <none>            142.74 146.74
## + Presion     3   140.43 150.43
## - Edad        1   154.56 156.56
##
## Step:  AIC=145.45
## Coronarios ~ Edad + IMC
##
##              Df Deviance    AIC
## + Colesterol  1   134.23 144.23
## <none>            137.45 145.45
## - IMC         2   142.74 146.74
## + Presion     3   134.78 148.78
## - Edad        1   146.50 152.50
##
## Step:  AIC=144.23
## Coronarios ~ Edad + IMC + Colesterol
##
##              Df Deviance    AIC
## <none>            134.23 144.23
## - Colesterol  1   137.45 145.45
## - IMC         2   139.93 145.93
## + Presion     3   131.28 147.28
## - Edad        1   139.29 147.29
```

Como puede verse, la selección stepwise finaliza con dos varables cuantitiavas seleccionadas (Edad y Colesterol) y una cualitativa (IMC). Obsérvese cómo el procedimiento de selección stepwise trata a cada variables cualitativa como una única variable, independientemente del número variables de diseño que necesite y de la codificación.

El modelo ajustado que se obtiene es el siguiente:

```
summary(Ajuste.Step)

##
## Call:
## glm(formula = Coronarios ~ Edad + IMC + Colesterol, family = binomial,
```

```
##     data = Chapman.Cuali)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1451  -0.5373  -0.3676  -0.2426   2.6808
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.536239   1.355164  -4.823 1.41e-06 ***
## Edad          0.046350   0.020925   2.215   0.0268 *
## IMCObesidad   1.614104   0.807307   1.999   0.0456 *
## IMCSobrepeso  0.914493   0.484120   1.889   0.0589 .
## Colesterol    0.006518   0.003614   1.803   0.0713 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 154.55  on 199  degrees of freedom
## Residual deviance: 134.23  on 195  degrees of freedom
## AIC: 144.23
##
## Number of Fisher Scoring iterations: 5
```

Un primer análisis del modelo resultante nos lleva a que todos los parámetros son significativos al 10% y dos de ellos incluso al 5%. La variable cualitativa incluída en el modelo tiene tres cateorías y por tanto dos variables de diseño y dos parámetros, ambos significativos, uno al 5% y el otro 10%. La categoría de referencia era *Normal* y las variables de diseño se corresponden con las categorías *Sobrepeso* y *Obesidad*.

La ecuación de predicción de este modelo será:

$$\begin{aligned}\hat{p} &= \frac{exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\tau}_1^P X_{21}^P + \hat{\tau}_2^P X_{22}^P + \hat{\beta}_3 X_3)}{1 + exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\tau}_1^P X_{21}^P + \hat{\tau}_2^P X_{22}^P + \hat{\beta}_3 X_3)} = \\ &= \frac{exp(-6.54 + (0.05)X_1 + (1.61)X_{21}^P + (0.91)X_{22}^P + (0.01)X_3)}{1 + exp(-6.54 + (0.05)X_1 + (1.61)X_{21}^P + (0.91)X_{22}^P + (0.01)X_3)}\end{aligned}$$

siendo X_{21}^P, X_{21}^P las variables de diseño con codificación parcial de la variable cualitativa *IMC*, $\hat{\tau}_1^P, \hat{\tau}_2^P$ los parámetros estimados asociados a tales variables de diseño, X_1, X_3 las variables cuantitativas (Edad y Colesterol) y $\hat{\beta}_1, \hat{\beta}_3$ los parámetros estimados asociados a tales variables.

El ajuste del modelo nos lleva a la siguiente interpretación de los parámetros:

- La exponencial del parámetro asociado a la variable de diseño asociada con la categoría *Obesidad de IMC* es el cociente de ventajas

$$\hat{\theta}_{21} = \exp(\hat{\tau}_1^P) = e^{1.6141038} = 5.0233841.$$

  Esto significa que la ventaja de padecer problemas coronarios (frente a no padecerlos) si tienes obesidad es 5.0233841 veces la ventaja cuando tienes IMC Normal, permaneciendo el resto de variables constantes. Dicho de otro modo, la ventaja se multiplica por 5.0233841 cuando se pasa de peso normal a obesidad, y permaneciendo el resto de variables constantes.

- La exponencial del parámetro asociado a la variable de diseño asociada con la categoría *sobrepeso de IMC* es el cociente de ventajas

$$\hat{\theta}_{31} = \exp(\hat{\tau}_2^P) = e^{0.9144932} = 2.4955101.$$

  Esto significa que la ventaja de padecer problemas coronarios (frente a no padecerlos) si tienes sobrepeso es 2.4955101 veces la ventaja cuando tienes peso normal, permaneciendo el resto de variables constantes. Dicho de otro modo, la ventaja se multiplica por 2.4955101 cuando se pasa de peso normal a sobrepeso, y permaneciendo el resto de variables constantes.

- La exponencial del parámetro asociado a la variable Edad es el cociente de ventajas

$$\hat{\theta}_1 = \exp(\hat{\beta}_1) = e^{0.0463495} = 1.0474404.$$

  Esto significa que la ventaja de padecer problemas coronarios (frente a no padecerlos) se multiplica por 1.0474404 cuando aumenta un año la edad, permaneciendo el resto de variables constantes

- La exponencial del parámetro asociado a la variable Colesterol es el cociente de ventajas

$$\hat{\theta}_3 = \exp(\hat{\beta}_3) = e^{0.0065181} = 1.0065394.$$

  Esto significa que la ventaja de padecer problemas coronarios (frente a no padecerlos) se multiplica por 1.0065394 cuando aumenta una unidad el nivel de colesterol, permaneciendo el resto de variables constantes

- La exponencial del parámetro independiente es la ventaja

$$\frac{\hat{p}_0}{1-\hat{p}_0} = exp\widehat{\beta}_0 = e^{-6.5362388} = 0.0014499.$$

  Esto significa que la ventaja de padecer problemas coronarios (frente a no padecerlos) para una persona de cero años, cero nivel de colesterol y perso normal es 0.0014499. Como se puede apreciar esta interpretación no tiene sentido.

Si en lugar de utilizar codificación parcial para las variables de diseño del IMC, se utiliza codificación marginal, utilizando también la primera categoría como categoría de referencia, el ajuste que se obtiene es diferente, los parámetros se interpretan de manera diferente, pero las conclusiones sobre el cambio en las ventajas son las mismas. La selección stepwise sería la misma, pues como se vio en el análisis de los test condicionales de razón de verosimilitudes, el tipo de codificación no afecta a estos tests.

1.4.1 Otros resultados del ajuste

Predicción

En el caso del ajuste múltiple, dado que se incluyen variables cuantitativas y cualitativas, es difícil encontrar muchos casos repetidos, por lo que los valores predichos por el modelo, que dependen de las observaciones de las variables explicativas, serán prácticamente tantos como observaciones. Los 10 primeros valores ajustados son

```
fitted.values(Ajuste.Step)[1:10]
```

```
##          1          2          3          4          5          6
## 0.12710808 0.08052962 0.12978394 0.03738917 0.32273538 0.23777720
##          7          8          9         10
## 0.07212666 0.39399978 0.10856742 0.12680924
```

Todo lo explicado en cuanto a probabilidades predichas con ajustes de regresión logística múltiple visto en el Tema 2, es aplicable aquí.

En cuanto a la función `predict()` para las observaciones de la muestra no hay ningún cambio con respecto a lo visto en el Tema 2. Sin embargo para observaciones nuevas es posible tal y como se vio en el tema 2. La única precaución que se debe tener es que el data.frame con el nuevo conjunto de datos tenga los nombres correctos de las variables (como se vio en el tema

2) y que las categorías de la variable cualitativa estén escritas exactamente como en el data.frame original. Véase por ejemplo el caso siguiente:

```
Nuevo.Chapman<-data.frame(c(66,66,66),c(200,220,300),
                          c("Normal","Sobrepeso","Normal"))
names(Nuevo.Chapman)<-c("Edad","Colesterol","IMC")
predict(Ajuste.Step,type="response",newdata = Nuevo.Chapman)

##         1         2         3
## 0.1021506 0.2444011 0.1792045
```

Como ejercicio, repita las sentencias anteriores cambiando por ejemplo la palabra *Normal* por la palabra *Nomal*. Verá cómo se produce un error.

Bondad del ajuste

En el caso de regresión logística múltiple en el que hay variables explicativas cualitativas, se puede estudiar la bondad del ajuste mediante el test de Hosmer y Lemeshow como en cualquier otro ajuste múltiple pues el número de probabilidades predicas diferentes es suficientemente grande.

```
##           Hosmer-Lemeshow Test
## X-squared           12.2130651
## p.value              0.1419476
```

El resultado del test indica que el ajuste logístico es adecuado para estos datos.

Significación de parámetros

El análisis de la significación de parámetros se lleva a cabo de igual modo que en cualquier otro ajuste de regresión logística. De hecho ya se hicieron los comentarios oportunos sobre este hecho al principio de la sección.

```
summary(Ajuste.Step)

##
## Call:
## glm(formula = Coronarios ~ Edad + IMC + Colesterol, family = binomial,
##     data = Chapman.Cuali)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1451  -0.5373  -0.3676  -0.2426   2.6808
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.536239   1.355164  -4.823 1.41e-06 ***
## Edad          0.046350   0.020925   2.215   0.0268 *
## IMCObesidad   1.614104   0.807307   1.999   0.0456 *
## IMCSobrepeso  0.914493   0.484120   1.889   0.0589 .
## Colesterol    0.006518   0.003614   1.803   0.0713 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 154.55  on 199  degrees of freedom
## Residual deviance: 134.23  on 195  degrees of freedom
## AIC: 144.23
##
## Number of Fisher Scoring iterations: 5
```

De los intervalos de confianza de los parámetros se puede hacer la misma reflexión que de la significación de los parámetros.

```
confint.default(Ajuste.Step)
```

```
##                      2.5 %      97.5 %
## (Intercept)  -9.1923109360 -3.88016662
## Edad          0.0053376116  0.08736139
## IMCObesidad   0.0318111485  3.19639650
## IMCSobrepeso -0.0343642557  1.86335059
## Colesterol   -0.0005655078  0.01360168
```

Como la exponencial de los parámetros tienen interpretación en términos de cocientes de ventajas para la codificación parcial, se pueden analizar los intervalos de confianza para los cocientes de ventaja mediate la exponencial de los intervalos de los parámetros.

```
exp(confint.default(Ajuste.Step))
```

```
##                      2.5 %      97.5 %
## (Intercept)  0.0001018193  0.02064738
## Edad         1.0053518820  1.09129100
## IMCObesidad  1.0323225313 24.44428639
## IMCSobrepeso 0.9662194896  6.44529615
## Colesterol   0.9994346521  1.01369460
```

Validación

El análisis de los residuos no tiene nada nuevo con lo visto en el Tema 2 (sólo se muestran aquí los 10 primeros):

```
residuals(Ajuste.Step,type= "pearson")[1:10]
```

```
##          1          2          3          4          5          6
## -0.3815982 -0.2959436 -0.3861864 -0.1970823  1.4486248 -0.5585270
##          7          8          9         10
## -0.2788069 -0.8063277 -0.3489840 -0.3810841
```

```
residuals(Ajuste.Step,type= "deviance")[1:10]
```

```
##          1          2          3          4          5          6
## -0.5214279 -0.4097742 -0.5272831 -0.2760655  1.5039432 -0.7369076
##          7          8          9         10
## -0.3869368 -1.0008745 -0.4794277 -0.5207710
```

Para detectar valores mál ajustados, los residuos estandarizados

```
rstandard(Ajuste.Step,type= "pearson")[1:10]
```

```
##          1          2          3          4          5          6
## -0.3847310 -0.2983756 -0.3890418 -0.1979720  1.4738115 -0.5714315
##          7          8          9         10
## -0.2801836 -0.8220776 -0.3515733 -0.3843917
```

```
rstandard(Ajuste.Step,type= "deviance")[1:10]
```

```
##          1          2          3          4          5          6
## -0.5257086 -0.4131416 -0.5311819 -0.2773118  1.5300917 -0.7539335
##          7          8          9         10
## -0.3888474 -1.0204245 -0.4829849 -0.5252910
```

y para valores influyentes, los valores hat y distancias de Cook

```
hatvalues(Ajuste.Step)[1:10]

##           1           2           3           4           5           6
## 0.016219218 0.016235103 0.014625574 0.008968480 0.033887027 0.044655725
##           7           8           9          10
## 0.009802727 0.037950168 0.014675699 0.017135272

cooks.distance(Ajuste.Step)[1:10]

##            1            2            3            4            5
## 4.880630e-04 2.938464e-04 4.492977e-04 7.093637e-05 1.523770e-02
##            6            7            8            9           10
## 3.052640e-03 1.554321e-04 5.331774e-03 3.681980e-04 5.151993e-04
```

Tabla de clasificación

Al igual que con la validación el análisis de la clasificación no tiene ninguna novedad. Se muestran a continuación los resultados:

En nuestro ejemplo la tabla de clasificación ha sido

	Clasificación	
Observación	Fracaso	éxito
Fracaso	VN=122	FP=52
éxito	FN=8	VP=18

Y las medidas:

- Tasa de clasificaciones correctas: CCR =70%
- Sensibilidad: 69.23%
- Especificidad: 70.11%

Curva ROC

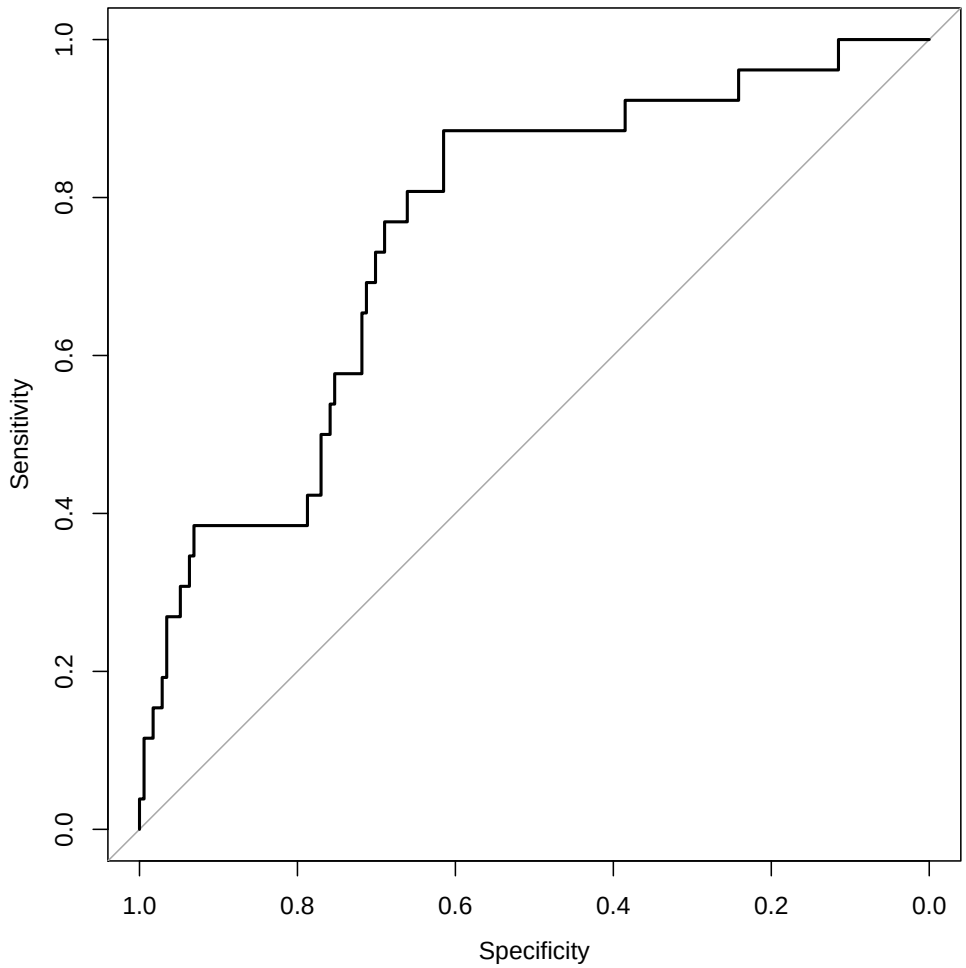
La Curva ROC tampoco presenta ninguna cuestión a comentar:

```
library(pROC)
CurvaROC.Step<-roc(Chapman.Cuali$Coronarios,fitted.values(Ajuste.Step))
CurvaROC.Step
```

```
##
## Call:
## roc.default(response = Chapman.Cuali$Coronarios, predictor = fitted.values(Ajuste
##
## Data: fitted.values(Ajuste.Step) in 174 controls (Chapman.Cuali$Coronarios 0) < 2
## Area under the curve: 0.7577
```

```
#CurvaROC.Step$thresholds
#CurvaROC.Step$sensitivities
#CurvaROC.Step$specificities
```

```
plot(CurvaROC.Step)
```



1.5 Análisis de datos agrupados

Cuando todas las variables disponibles para un análisis de regresión logística son cualitativas, es habitual que la información disponible se presente en una tabla de frecuencias multidimensional.

Veamos el siguiente ejemplo: Para estudiar si el desarrollo de una enfermedad coronaria (EC) está influenciado por el número de cigarrillos diarios que fuma una persona, su estilo de vida (A y B) y su presión sanguínea (PS), se han tomado 16 muestras independientes de individuos, una para cada una de las posibles combinaciones de niveles de las tres variables explicativas. La información se muestra a continuación.

			N^0 cigarrillos			
			0	1-20	21-30	> 30
PS ≥ 140	Tipo A	EC	29	21	7	12
		No EC	155	76	45	43
PS ≥ 140	Tipo B	EC	8	9	3	7
		No EC	171	62	31	14
PS < 140	Tipo A	EC	41	24	27	17
		No Ec	599	277	140	116
PS < 140	Tipo B	EC	20	16	13	3
		No EC	669	320	139	80

Para la lectura de datos es conveniente formatear el data.frame convenientemente. Según se desprende en el enunciado, se tomaron 16 muestras independiente de individuos, para un total de 3194 a los que se les midieron un conjunto de variables. Todas las variables bajo estudio son variables cualitativas, la variable *Número de cigarrillos* tiene cuatro categorías {0, 1-20, 21-30 y >30}, la variable *presión sanguínea* dos categorías {$\geq$ 140 y <140}, la variable *tipo de vida* otras dos categorías {A y B} y la variable respuesta *enfermedad coronária* otras dos categorías {Enfermo y No enfermo}. La información de las variables está disponible a través de una tabla de frecuencias que indica el número de individuos que resultaron en cada combinación de categorías de las variables. Este es un caso de datos agrupados.

La lectura de un conjunto de datos como este en R, se puede hacer de diversas formas. Quizás el método más sencillo sea crear un fichero de texto con las distintas combinaciones de categorías de las distintas variables y el número de éxitos y fracasos en cada combinación. Éste es el caso se muestra a continuación:

```
-----------------------------------------------------------------------
Ciga Compor Psang EC1 EC0
C1 A Alta 29 155
```

```
C2 A Alta 21 76
C3 A Alta 7 45
C4 A Alta 12 43
C1 B Alta 8 171
C2 B Alta 9 62
C3 B Alta 3 31
C4 B Alta 7 14
C1 A Baja 41 599
C2 A Baja 24 277
C3 A Baja 27 140
C4 A Baja 17 116
C1 B Baja 20 669
C2 B Baja 16 320
C3 B Baja 13 139
C4 B Baja 3 80
```

Cada variable ha de repetir sus categorías de tal manera que se obtengan todas las posibles combinaciones posibles de categorías de las variables disponibles. Además se crean dos columnas (variables) destinadas a contener la frecuencia de éxitos y fracasos (enfermos y no enfermos) en dichas combinaciones de categorías. Para la configuración de los datos se tiene en cuenta lo siguiente: la variable *Número de cigarrillos* ha sido codificada como {C1, C2, C3 y C4} respectivamente para las caregorías {0, 1-20, 21-30 y >30}, la variable *presión sanguínea* como {Alta y Baja} respectivamente para {$\geq$ 140 y <140}, la variable *tipo de vida* como {A y B}. Las columnas EC1 y EC0 contienen el número de enfermos y no enfermos respectivamente. Esta manera de codificar los datos, hace que la manera que tiene R de considerar las variables de tipo `factor` sea la que se espera al tomar por defecto el orden alfabético.

Con esta codificación ya se pueden leer los datos y ajustar los modelos simples o múltiples que se desee. Observe que al tratarse de datos agrupados, en la función `glm()` se debe incluir la respuesta con el número de éxitos y de fracasos a través de la sentencia `cbind()`

```
Coronarios<-read.table("Coronarios.txt",header=T,sep="")
Ajuste.Coronarios0<-glm(cbind(EC1,EC0)~1,family=binomial,
                        data=Coronarios)
Ajuste.Coronarios.All<-glm(cbind(EC1,EC0)~Ciga+Compor+Psang,family=binomial,
                           data=Coronarios)
Ajuste.Coronarios.Step<-step(Ajuste.Coronarios0,scope=list(lower=cbind(EC1,EC0)~1,
upper=cbind(EC1,EC0)~Ciga+Compor+Psang),direction = "both")
```

```
## Start:  AIC=185.55
## cbind(EC1, EC0) ~ 1
##
##          Df Deviance    AIC
## + Compor  1   77.110 149.27
## + Psang   1   79.615 151.78
## + Ciga    3   83.913 160.07
## <none>       115.385 185.55
##
## Step:  AIC=149.27
## cbind(EC1, EC0) ~ Compor
##
##          Df Deviance    AIC
## + Psang   1   44.802 118.96
## + Ciga    3   50.209 128.37
## <none>        77.110 149.27
## - Compor  1  115.385 185.55
##
## Step:  AIC=118.96
## cbind(EC1, EC0) ~ Compor + Psang
##
##          Df Deviance     AIC
## + Ciga    3   18.766  98.927
## <none>        44.802 118.964
## - Psang   1   77.110 149.272
## - Compor  1   79.615 151.777
##
## Step:  AIC=98.93
## cbind(EC1, EC0) ~ Compor + Psang + Ciga
##
##          Df Deviance     AIC
## <none>        18.766  98.927
## - Ciga    3   44.802 118.964
## - Compor  1   48.943 127.104
## - Psang   1   50.209 128.370
```

Otra forma de realizar la carga de datos de este ejercicio es a través tablas de frecuenicas multidimensionales. En R es posible crear tablas de frecuencias multidimensionales con la sentencia `array()`, que contiene por un lado el vector de frecuencias y por otro las dimensiones de cada variable. Además se puede añadir nombres a las variables y las categorías con la sentencia

`dimnames()`. A continuación se puede ver cómo crear la tabla de frecuencias de este ejercicio.

```
## , , Tipo = A, Presion = Alta
##
##     Ciga
## Ec    C1 C2 C3 C4
##   Si  29 21  7 12
##   No 155 76 45 43
##
## , , Tipo = B, Presion = Alta
##
##     Ciga
## Ec    C1 C2 C3 C4
##   Si   8  9  3  7
##   No 171 62 31 14
##
## , , Tipo = A, Presion = Baja
##
##     Ciga
## Ec    C1  C2  C3  C4
##   Si  41  24  27  17
##   No 599 277 140 116
##
## , , Tipo = B, Presion = Baja
##
##     Ciga
## Ec    C1  C2  C3 C4
##   Si  20  16  13  3
##   No 669 320 139 80
```

Finalmente, a partir de la tabla de frecuencias se puede obtener el data.frame, necesario para el ajuste del siguiente modo:

```
Coronarios.Tabla<-as.data.frame(as.table(TablaFuma))
Coronarios.Tabla

##    Ec Ciga Tipo Presion Freq
## 1  Si   C1    A    Alta   29
## 2  No   C1    A    Alta  155
## 3  Si   C2    A    Alta   21
```

```
## 4  No   C2    A    Alta   76
## 5  Si   C3    A    Alta    7
## 6  No   C3    A    Alta   45
## 7  Si   C4    A    Alta   12
## 8  No   C4    A    Alta   43
## 9  Si   C1    B    Alta    8
## 10 No   C1    B    Alta  171
## 11 Si   C2    B    Alta    9
## 12 No   C2    B    Alta   62
## 13 Si   C3    B    Alta    3
## 14 No   C3    B    Alta   31
## 15 Si   C4    B    Alta    7
## 16 No   C4    B    Alta   14
## 17 Si   C1    A    Baja   41
## 18 No   C1    A    Baja  599
## 19 Si   C2    A    Baja   24
## 20 No   C2    A    Baja  277
## 21 Si   C3    A    Baja   27
## 22 No   C3    A    Baja  140
## 23 Si   C4    A    Baja   17
## 24 No   C4    A    Baja  116
## 25 Si   C1    B    Baja   20
## 26 No   C1    B    Baja  669
## 27 Si   C2    B    Baja   16
## 28 No   C2    B    Baja  320
## 29 Si   C3    B    Baja   13
## 30 No   C3    B    Baja  139
## 31 Si   C4    B    Baja    3
## 32 No   C4    B    Baja   80
```

Y se puede ajustar el modelo logístico de este modo

```
Ajuste.Coronarios.Tabla<-glm(Ec~Ciga+Tipo+Presion,weights = Freq,data=Coronar
Ajuste.Coronarios.Tabla

##
## Call:  glm(formula = Ec ~ Ciga + Tipo + Presion, family = binomial,
##     data = Coronarios.Tabla, weights = Freq)
##
## Coefficients:
```

```
## (Intercept)       CigaC2       CigaC3       CigaC4        TipoB
##      1.8863      -0.4394      -0.8067      -0.7795       0.7566
## PresionBaja
##      0.8048
##
## Degrees of Freedom: 31 Total (i.e. Null);  26 Residual
## Null Deviance:     1788
## Residual Deviance: 1691  AIC: 1703
```