Departamento de Estadística e I.O.

Máster en Estadística Aplicada

UNIVERSIDAD
DE GRANADA

MODELOS DE RESPUESTA DISCRETA
APLICACIONES BIOSANITARIAS

Tema 4 de prácticas
Ajuste de regresión logit multinomial con R

Profesores

Ana María Aguilera del Pino

Manuel Escabias Machuca

Título original: Modelos de Respuesta Discreta. Aplicaciones Biosanitarias.
Tema 4 de prácticas: Ajuste de regresión logit multinomial con R

Índice general

Capítulo 1

Ajuste de regresión logit de respuesta multinomial con R

1.1. Introducción

Este capítulo tiene por objetivo mostrar el ajuste de modelos de regresión logit de respuesta múltiple tanto con variables explicativas cualitativas como cuantitativas. Para ilustrar todas las cuestiones del tratamiento de variables cualitativas en regresión logística, se utilizará el mismo conjunto de datos del Tema 3. 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.

Recuérdese que para leer este conjunto de datos, se recurre a la sentencia `read.csv()` (si se utiliza 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).

Recuérdese además que 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 el orden sea *Optima, Normal, Alta, Descompensada* y que para el IMC queremos que sea *Normal, Sobrepeso, Obesidad*, las sentencias serían

```
Chapman.Cuali$Presion<-factor(Chapman.Cuali$Presion,
levels=c("Optima","Normal","Alta","Descompensada"))

Chapman.Cuali$IMC<-factor(Chapman.Cuali$IMC,
levels=c("Normal","Sobrepeso","Obesidad"))
```

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)

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

contrasts(Chapman.Cuali$IMC)

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

Nota: en los apuntes del tema 3 inicialmente se indicó erróneamente que la función `levels()<-` reordenaba los niveles, cuando lo que hacía era reordenar los valores de la variable. Tras ese error, se corrigieron los apuntes y la relación de ejercicios avisando convenientemente, utilizando la sentencia `relevel(,ref=)` que sólo cambiaba la categoría de referencia. Ahora con la sentencia propuesta, se reordenan todos los niveles de una variable de tipo factor, manteniendo las observaciones en el conjunto de datos.

De ahora en adelante se asumirá estas categorías de referencia y estos órdenes para las variables cualitativas de este ejemplo, en lugar del que asigna por defecto R, esto es, asumimos que hemos cambiado el orden.

1.2. Analisis de regresión logit de respuesta nominal con R

1.2.1. Ajuste de regresión logit de respuesta nominal con variables explicativas cuantitativas y observaciones en un *Data.Frame*

Supongamos que se quiere modelizar la presión arterial en función de la edad y el nivel de colesterol (ambas cuantitativas). En este caso la variable respuesta Y (presión arterial) toma cuatro valores que denotaremos por Y_0 =Optima (categoría de referencia), Y_1 =Normal, Y_2 =Alta, Y_3 =Descompensada (S=4). Las variables explicativas cuantitativas son X_1 =Edad y X_2 =Colesterol

R no posee ninguna función estándar para realizar este ajuste. Sin embargo existen varias librerías que incluyen funciones para realizar estos ajustes.

La función que se explica a continuación pertenece a la librería `nnet` y la función es `multinom`.

```
library(nnet)
Ajuste.Multinom.Cuanti<-multinom(Presion~Edad+Colesterol,
data=Chapman.Cuali)

## # weights:  16 (9 variable)
## initial  value 277.258872
## iter  10 value 218.912397
## final  value 218.133323
## converged

summary(Ajuste.Multinom.Cuanti)

## Call:
## multinom(formula = Presion ~ Edad + Colesterol, data = Chapman.Cuali)
##
## Coefficients:
##                (Intercept)         Edad   Colesterol
## Normal          -0.8852911 -0.002834799  0.001237950
## Alta            -8.5771654  0.165655037 -0.001811441
## Descompensada   -2.0816026  0.039881594  0.003566166
##
## Std. Errors:
##                (Intercept)       Edad  Colesterol
## Normal           1.1235970 0.02476033 0.003977586
## Alta             2.0727371 0.03534176 0.005548131
## Descompensada    0.8786556 0.01794607 0.002979385
##
## Residual Deviance: 436.2666
## AIC: 454.2666
```

El resultado del ajuste con la función `summary()` muestra los parámetros estimados y sus errores estándar.

La función `multinom()` toma las transformaciones logit generalizadas con respecto a la categoría de referencia. Para el ejemplo que nos ocupa, y adaptando la notación a la salida que proporciona la función,

$$L_s(x) = \ln\left[\frac{p_s(x)}{p_0(x)}\right], \ \forall s = 1, 2, 3.$$

donde $L_s(x)$ representa el logaritmo de la ventaja de respuesta Y_s frente a la respuesta Y_1. La modelización en términos de las variables explicativas será

$$L_s(x) = \beta_{s0} + \beta_{s1}X_1 + \beta_{s2}X_2, \ \forall s = 1, 2, 3.$$

De manera desarrollada:

$$\begin{aligned} L_1(x) &= \beta_{10} + \beta_{11}X_1 + \beta_{12}X_2 \\ L_2(x) &= \beta_{20} + \beta_{21}X_1 + \beta_{22}X_2 \\ L_3(x) &= \beta_{30} + \beta_{31}X_1 + \beta_{32}X_2 \end{aligned}$$

Obsérvese, que para cada variable explicativa se tienen tres parámetros (número de categorías de la respuesta menos uno): para $X_1 : \beta_{11}, \beta_{21}, \beta_{31}$ y para $X_2 : \beta_{12}, \beta_{22}, \beta_{32}$. También hay tres parámetros independientes: $\beta_{10}, \beta_{20}, \beta_{30}$.
Los parámetros estimados del ajuste son:

$\widehat{\beta}_{10}$ =-0.8852911	$\widehat{\beta}_{11}$ =-0.0028348	$\widehat{\beta}_{12}$ =0.0012379
$\widehat{\beta}_{20}$ =-8.5771654	$\widehat{\beta}_{21}$ =0.165655	$\widehat{\beta}_{22}$ =-0.0018114
$\widehat{\beta}_{30}$ =-2.0816026	$\widehat{\beta}_{31}$ =0.0398816	$\widehat{\beta}_{32}$ =0.0035662

Al igual que en el resto de modelos logit, la exponencial de los parámetros estimados se pueden interpretar en términos de cocientes de ventajas:

```
exp(summary(Ajuste.Multinom.Cuanti)$coefficients)

##                  (Intercept)      Edad Colesterol
## Normal          0.4125940545 0.9971692  1.0012387
## Alta            0.0001883581 1.1801659  0.9981902
## Descompensada 0.1247301538 1.0406875  1.0035725
```

- $\exp(\widehat{\beta}_{10})$ =0.4125941 es la ventaja de presión normal frente a óptima para individuos con Edad=0 y Colesterol=0. Esta interpretacón no tiene sentido pues en la muestra no hay individuos con Edad=0 y Colesterol=0

- $\exp(\widehat{\beta}_{20})$ =$1{,}8835813 \times 10^{-4}$ es la ventaja de presión alta frente a óptima para individuos con Edad=0 y Colesterol=0. Esta interpretacón no tiene sentido pues en la muestra no hay individuos con Edad=0 y Colesterol=0

- $\exp(\widehat{\beta}_{30})$ =0.1247302 es la ventaja de presión descompensada frente a óptima para individuos con Edad=0 y Colesterol=0. Esta interpretacón no tiene sentido pues en la muestra no hay individuos con Edad=0 y Colesterol=0.

- $\exp(\widehat{\beta}_{11})$ =0.9971692 es el cociente de ventajas de presión normal frente a óptima cuando un individuo aumenta su edad en 1 año permaneciendo constante el nivel de colesterol. Equivalentemente, sería el cambio multiplicativo que se produce en la ventaja de presión normal frente a óptima, cuando se aumenta un año la edad, permaneciendo constante el nivel de colesterol.

- $\exp(\widehat{\beta}_{12})$ =1.0012387 es el cociente de ventajas de presión normal frente a óptima cuando un individuo aumenta su nivel de colesterol en 1 unidad permaneciendo constante la edad. Equivalentemente, sería el cambio multiplicativo que se produce en la ventaja de presión normal frente a óptima, cuando se aumenta el nivel de colesterol una unidad, permaneciendo constante la edad.

- $\exp(\widehat{\beta}_{21})$ =1.1801659 es el cociente de ventajas de presión alta frente a óptima cuando un individuo aumenta su edad en 1 año permaneciendo constante el nivel de colesterol. Equivalentemente, sería el cambio multiplicativo que se produce en la ventaja de presión alta frente a óptima, cuando se aumenta un año la edad, permaneciendo constante el nivel de colesterol.

- $\exp(\widehat{\beta}_{22})$ =0.9981902 es el cociente de ventajas de presión alta frente a óptima cuando un individuo aumenta su nivel de colesterol en 1 unidad permaneciendo constante la edad. Equivalentemente, sería el cambio multiplicativo que se produce en la ventaja de presión alta frente a óptima, cuando se aumenta el nivel de colesterol una unidad, permaneciendo constante la edad.

- $\exp(\widehat{\beta}_{31})$ =1.0406875 es el cociente de ventajas de presión descompensada frente a óptima cuando un individuo aumenta su edad en 1 año permaneciendo constante el nivel de colesterol. Equivalentemente, sería el cambio multiplicativo que se produce en la ventaja de presión descompensada frente a óptima, cuando se aumenta un año la edad, permaneciendo constante el nivel de colesterol.

- $\exp(\widehat{\beta}_{32})$ =1.0035725 es el cociente de ventajas de presión descompensada frente a óptima cuando un individuo aumenta su nivel de colesterol en 1 unidad permaneciendo constante la edad. Equivalentemente, sería el cambio multiplicativo que se produce en la ventaja de presión descompensada frente a óptima, cuando se aumenta el nivel de colesterol una unidad, permaneciendo constante la edad.

Como siempre, la estimación de los parámetros se ha interpretado sin tener en cuenta la significación de estos parámetros. La función `multinom` no muestra por defecto la significación de parámetros. Para ello, asumiendo la distribución normal asintótica de los parámetros (al igual que en regresión logística) se pueden obtener los valores experimentales del test de Wald para cada parámetro a partir de la sentencia `summary()`

```
summary(Ajuste.Multinom.Cuanti)$coefficients/
summary(Ajuste.Multinom.Cuanti)$standard.errors

##               (Intercept)       Edad Colesterol
## Normal         -0.7879081 -0.1144895  0.3112315
## Alta           -4.1380866  4.6872329 -0.3264957
## Descompensada  -2.3690769  2.2223021  1.1969472
```

Así mismo se pueden obtener los p-valores con las probabilidades de la distribución normal:

```
2*pnorm(abs(summary(Ajuste.Multinom.Cuanti)$coefficients/
summary(Ajuste.Multinom.Cuanti)$standard.errors),lower.tail=F)

##                (Intercept)         Edad Colesterol
## Normal        4.307505e-01 9.088498e-01  0.7556247
## Alta          3.502142e-05 2.769238e-06  0.7440493
## Descompensada 1.783255e-02 2.626290e-02  0.2313271
```

También se podrían utilizar los intervalos de confianza de los parámetros para estudiar su significación:

```
confint(Ajuste.Multinom.Cuanti)

## , , Normal
##
##                    2.5 %      97.5 %
## (Intercept) -3.087500645 1.316918468
## Edad        -0.051364159 0.045694562
## Colesterol  -0.006557975 0.009033875
```

```
##
## , , Alta
##
##                     2.5 %       97.5 %
## (Intercept) -12.63965543 -4.514675439
## Edad          0.09638647  0.234923605
## Colesterol   -0.01268558  0.009062696
##
## , , Descompensada
##
##                     2.5 %       97.5 %
## (Intercept) -3.803735975 -0.359469315
## Edad         0.004707939  0.075055250
## Colesterol  -0.002273321  0.009405653
```

Así mismo se podría estudiar la significación de los cocientes de ventajas, mediante los intervalos de confianza:

```
exp(confint(Ajuste.Multinom.Cuanti))

## , , Normal
##
##                   2.5 %   97.5 %
## (Intercept) 0.04561582 3.731904
## Edad        0.94993268 1.046755
## Colesterol  0.99346348 1.009075
##
## , , Alta
##
##                     2.5 %     97.5 %
## (Intercept) 3.240913e-06 0.01094716
## Edad        1.101185e+00 1.26481214
## Colesterol  9.873945e-01 1.00910389
##
## , , Descompensada
##
##                   2.5 %    97.5 %
## (Intercept) 0.02228735 0.6980467
## Edad        1.00471904 1.0779437
## Colesterol  0.99772926 1.0094500
```

A partir de la estimación de parámetros, se pueden obtener las estimaciones de la probabilidades de cada categoría de la respuesta. Obsérvese que las probabilidades predichas por el modelo se pueden obtener del siguiente modo para observaciones x_1 de la variable X_1 y x_2 de la variable X_2:

$$
\begin{aligned}
p_s(x) &= \frac{\exp\left(\widehat{\beta}_{s0} + \widehat{\beta}_{s1}x_1 + \widehat{\beta}_{s2}x_2\right)}{1 + \sum_{s=1}^{3} \exp\left(\widehat{\beta}_{s0} + \widehat{\beta}_{s1}x_1 + \widehat{\beta}_{s2}x_2\right)} \quad \forall s = 1, 2, 3, \\
p_0(x) &= \frac{1}{1 + \sum_{s=1}^{3} \exp\left(\widehat{\beta}_{s0} + \widehat{\beta}_{s1}x_1 + \widehat{\beta}_{s2}x_2\right)}.
\end{aligned}
$$

Las probabilidades predichas de cada categoría de la respuesta, en cada valor/es de la variable/s explicativa/s las proporciona el objeto de tipo `multinom()` con la sentencia `predict()` y la opción `type="probs"`. Para las 10 primeras observaciones del fichero de datos son:

```
predict(Ajuste.Multinom.Cuanti,type="probs")[1:10,]

##        Optima     Normal        Alta Descompensada
## 1   0.2892624 0.14427931 0.050344242     0.5161141
## 2   0.3664964 0.18430417 0.014731729     0.4344677
## 3   0.2984441 0.15484935 0.030201198     0.5165054
## 4   0.3651516 0.19611876 0.006986669     0.4317430
## 5   0.1178427 0.06040616 0.306905135     0.5148460
## 6   0.1271008 0.06011467 0.372378854     0.4404057
## 7   0.2649157 0.13953261 0.042574620     0.5529771
## 8   0.1262782 0.07110450 0.176570578     0.6260467
## 9   0.1956264 0.10223466 0.115764354     0.5863746
## 10 0.1850391 0.09998855 0.104272541     0.6106998
```

Obsérvese cómo las probabilidades por filas suman la unidad como es de esperar en una distribución multinomial.

Una vez estimadas las probabilidades de cada categoría, el modelo predice, para cada valor/es de la variable/s explicativa/s la categoría de la respuesta con mayor probabilidad. Estas predicciones también las proporciona la función `multinom()` con la sentencia `predict()` con la opción `type=çlass"`. Para las 10 primeras observaciones del fichero de datos son:

```
predict(Ajuste.Multinom.Cuanti,type="class")[1:10]

##  [1] Descompensada Descompensada Descompensada Descompensada Descompensada
##  [6] Descompensada Descompensada Descompensada Descompensada Descompensada
## Levels: Optima Normal Alta Descompensada
```

Estas predicciones permiten obtener una tabla de clasificación que tiene por filas las observaciones y por columnas las predicciones:

```
table(Chapman.Cuali$Presion,
predict(Ajuste.Multinom.Cuanti,type="class"))

##
##                 Optima Normal Alta Descompensada
##   Optima            15      0    1            40
##   Normal             7      0    0            22
##   Alta               0      0    2            15
##   Descompensada     13      0    4            81
```

Obsérvese cómo con este modelo:

- Entre los individuos con presión óptima, el modelo acierta en un 26.7857143 por ciento.
- Entre los individuos con presión normal, el modelo acierta en un 0 por ciento
- Entre los individuos con presión alta, el modelo acierta en un 11.7647059 por ciento
- Entre los individuos con presión descompensada, el modelo acierta en un 82.6530612 por ciento
- Entre todos los casos, el modelo acierta en un 49 por ciento

El estudio de la bondad del ajuste del modelo multinomial sólo es posible realizarlo comparando con el modelo saturado, esto es el que tiene tantos parámetros estimados como observaciones. En el caso de datos no agrupados la log-verosimilitud del modelo saturado es nula, por tanto, para estudiar la bondad del ajuste del modelo multinomial en este caso el valor experimental del test (estadístico) será la *deviance* del modelo y el p-valor el de la Chi cuadrado correspondiente:

```
Ajuste.Multinom.Cuanti$deviance

## [1] 436.2666

pchisq(Ajuste.Multinom.Cuanti$deviance,591,lower.tail = F)

## [1] 0.9999996
```

Hay que tener en cuenta que el modelo saturado tiene 3 parámetros libres (probabilidades) en cada combinación diferente de observaciones de las variables explicativas. En nuestro ejemplo, se tienen 200 combinaciones diferentes de observaciones de las variables explicativas (tantas como observaciones o casos) por lo que el número de parámetros libres del modelo saturado es 600. Como el modelo ajustado tiene 9 parámetros, los grados de libertad de la distribución Chi-Cuadrado del test de bondad de ajuste serán 591. Con esto, aceptamos el modelo multinomial como modelo adecuado para estos datos, frente al saturado.

1.2.2. Ajuste de regresión logit de respuesta nominal con variables explicativas cuantitativas y cualitativas y observaciones en un *Data.Frame*

Supongamos que se quiere modelizar la presión arterial en función de la edad, el nivel de colesterol (ambas cuantitativas) y el IMC (cualitativa). En este ejemplo se mostrarán los mismos resultados vistos en la sección anterior, pero en este caso sin las interpretaciones. Sólo aquellas cuestiones que difieran en algún aspecto con respecto a lo anterior se explicarán con cierto detalle.

```
Ajuste.Multinom.Cuanli<-multinom(Presion~Edad+Colesterol+IMC,
data=Chapman.Cuali)

## # weights:  24 (15 variable)
## initial  value 277.258872
## iter  10 value 220.279576
## iter  20 value 212.060048
## iter  30 value 212.002261
## final  value 212.002125
## converged

summary(Ajuste.Multinom.Cuanli)
```

```
## Call:
## multinom(formula = Presion ~ Edad + Colesterol + IMC, data = Chapman.Cuali
##
## Coefficients:
##              (Intercept)         Edad   Colesterol IMCSobrepeso IMCObesid
## Normal         -1.002555 -0.009621724  0.001565181    0.7303200    11.081
## Alta           -8.513822  0.160453799 -0.002069114    0.5681986    10.750
## Descompensada  -2.234389  0.030146743  0.004078128    0.8769690    11.797
##
## Std. Errors:
##              (Intercept)       Edad  Colesterol IMCSobrepeso IMCObesidad
## Normal         1.1319925 0.02530219 0.004032726    0.4888471   0.8227964
## Alta           2.0822288 0.03611116 0.005778952    0.6571189   0.8853329
## Descompensada  0.8964059 0.01852490 0.003091552    0.3766979   0.5871605
##
## Residual Deviance: 424.0043
## AIC: 454.0043
```

El resultado del ajuste con la función `summary()` muestra los parámetros estimados y sus errores estándar.

En este caso en la modelización de las transformaciones logit generalizadas se incluyen además las variables de diseño de la variable cualitativa IMC:

$$L_s(x) = \beta_{s0} + \beta_{s1}X_1 + \beta_{s2}X_2 + \tau_{s1}X_{31} + \tau_{s2}X_{32},\ \forall s = 1, 2, 3$$

donde se ha denotado por X_{31} y X_{32} a las variables de diseño de la variable IMC.

Al igual que en el resto de modelos logit, la exponencial de los parámetros estimados se pueden interpretar en términos de cocientes de ventajas:

```
exp(summary(Ajuste.Multinom.Cuanli)$coefficients)
```

```
##                (Intercept)      Edad Colesterol IMCSobrepeso IMCObesidad
## Normal        0.3669406279 0.9904244   1.001566     2.075745    64985.67
## Alta          0.0002006755 1.1740435   0.997933     1.765085    46648.08
## Descompensada 0.1070575681 1.0306058   1.004086     2.403603   132950.21
```

Para los parámetros asociados a la variable IMC se tienen las siguientes interpretaciones:

- $\exp(\widehat{\beta}_{10})$ =0.3669406 es la ventaja de presión normal frente a óptima para individuos con Edad=0, Colesterol=0 e IMC=Normal. Esta interpretacón no tiene sentido pues en la muestra no hay individuos con

Edad=0, Colesterol=0 e IMC=Normal. Análogamente se interpretarían el resto de parámetros independientes.

- $\exp(\widehat{\tau}_{11})$ =2.0757447 es el cociente de ventajas de presión normal frente a óptima cuando un individuo tiene sobrepeso en lugar de peso normal permaneciendo constantes la edad y el nivel de colesterol. Equivalentemente, sería el cambio multiplicativo que se produce en la ventaja de presión normal frente a óptima, cuando se pasa de IMC Normal a Sobrepeso, permaneciendo constantes la edad y el nivel de colesterol. Análogamente se interpretarían para el resto de categorías de presión

- $\exp(\widehat{\tau}_{12})$ $=6{,}4985671 \times 10^4$ es el cociente de ventajas de presión normal frente a óptima cuando un individuo tiene Obesidad en lugar de peso normal permaneciendo constantes la edad y el nivel de colesterol. Equivalentemente, sería el cambio multiplicativo que se produce en la ventaja de presión normal frente a óptima, cuando se pasa de IMC Normal a Obesidad, permaneciendo constantes la edad y el nivel de colesterol. Análogamente se interpretarían para el resto de categorías de presión

Como ya se indicó en el caso de regresión logit binaria con variables explicativas cualitativas, la significación de la variable IMC no se puede estudiar a partir de la significación de los parámetros, sino a través de los test condicionales de razón de verosimilitudes. En presencia de las variables cuantitativas Edad y Colesterol, la significación de la variable IMC se estudiaría del siguiente modo:

```
anova(multinom(Presion~Edad+Colesterol+IMC,
data=Chapman.Cuali),multinom(Presion~Edad+Colesterol,
data=Chapman.Cuali))

## # weights:  24 (15 variable)
## initial  value 277.258872
## iter  10 value 220.279576
## iter  20 value 212.060048
## iter  30 value 212.002261
## final  value 212.002125
## converged
## # weights:  16 (9 variable)
## initial  value 277.258872
## iter  10 value 218.912397
## final  value 218.133323
```

```
## converged
##                        Model Resid. df Resid. Dev   Test    Df LR stat.     Pr
## 1        Edad + Colesterol       591   436.2666             NA        NA
## 2 Edad + Colesterol + IMC       585   424.0043 1 vs 2     6  12.2624 0.056
```

Obsérvese cómo el valor experimental del test es 12.2624 que para una Chi-cuadrado con 6 grados de libertad arroja un p-valor de 0.05636, que al 5 % de significación, me indica que la variable IMC no debe estar en el modelo (en presencia de la edad y el colesterol), pero no así a una significación del 10 %.

Los intervalos de confianza de los cocientes de ventajas:

```
exp(confint(Ajuste.Multinom.Cuanli))

## , , Normal
##
##                       2.5 %       97.5 %
## (Intercept)  3.990635e-02 3.374035e+00
## Edad         9.425059e-01 1.040779e+00
## Colesterol   9.936812e-01 1.009514e+00
## IMCSobrepeso 7.962813e-01 5.411048e+00
## IMCObesidad  1.295544e+04 3.259742e+05
##
## , , Alta
##
##                       2.5 %       97.5 %
## (Intercept)  3.389206e-06 1.188203e-02
## Edad         1.093821e+00 1.260149e+00
## Colesterol   9.866937e-01 1.009300e+00
## IMCSobrepeso 4.868833e-01 6.398912e+00
## IMCObesidad  8.226915e+03 2.645029e+05
##
## , , Descompensada
##
##                       2.5 %       97.5 %
## (Intercept)  1.847546e-02 6.203538e-01
## Edad         9.938575e-01 1.068713e+00
## Colesterol   9.980208e-01 1.010189e+00
## IMCSobrepeso 1.148727e+00 5.029313e+00
## IMCObesidad  4.206269e+04 4.202241e+05
```

Las probabilidades predichas y las categorías predichas (las 10 primeras):

```
predict(Ajuste.Multinom.Cuanli,type="probs")[1:10,]

##        Optima     Normal        Alta Descompensada
## 1  0.21196738 0.15733815 0.051687798     0.5790067
## 2  0.26191578 0.20740448 0.015513269     0.5151665
## 3  0.21272961 0.16901400 0.030438919     0.5878175
## 4  0.43313062 0.18396069 0.006984248     0.3759244
## 5  0.09233875 0.06402897 0.303612405     0.5400199
## 6  0.10127448 0.06343205 0.380929686     0.4543638
## 7  0.34854920 0.13352547 0.043962154     0.4739632
## 8  0.09402632 0.07476195 0.165626046     0.6655857
## 9  0.27600997 0.09975910 0.122582926     0.5016480
## 10 0.26136375 0.09854313 0.109771159     0.5303220

predict(Ajuste.Multinom.Cuanli,type="class")[1:10]

##  [1] Descompensada Descompensada Descompensada Optima        Descompensada
##  [6] Descompensada Descompensada Descompensada Descompensada Descompensada
## Levels: Optima Normal Alta Descompensada
```

La tabla de clasificación:

```
table(Chapman.Cuali$Presion,
predict(Ajuste.Multinom.Cuanli,type="class"))

##
##                 Optima Normal Alta Descompensada
##   Optima            24      0    1            31
##   Normal             5      0    0            24
##   Alta               0      0    2            15
##   Descompensada     13      0    3            82
```

La bondad del ajuste:

```
Ajuste.Multinom.Cuanli$deviance

## [1] 424.0043

pchisq(Ajuste.Multinom.Cuanli$deviance,585,lower.tail = F)

## [1] 0.9999999
```

Que indica que el modelo de respuesta nominal es adecuado.

Antes de terminar con este apartado vamos a mostrar el resultado de una selección stepwise de estas variables en un modelo multinomial:

```
Ajuste.Multinom.0<-multinom(Presion~1,data=Chapman.Cuali)

## # weights:  8 (3 variable)
## initial  value 277.258872
## final  value 239.100763
## converged

Ajuste.Multinom.Step<-step(Ajuste.Multinom.0,
scope=list(lower=Presion~1,upper=Presion~Edad+Colesterol+IMC),
direction="both")

## Start:  AIC=484.2
## Presion ~ 1
##
## trying + Edad
## # weights:  12 (6 variable)
## initial  value 277.258872
## iter  10 value 219.351589
## final  value 219.294302
## converged
## trying + Colesterol
## # weights:  12 (6 variable)
## initial  value 277.258872
## iter  10 value 235.202246
## final  value 235.202241
## converged
## trying + IMC
## # weights:  16 (9 variable)
## initial  value 277.258872
## iter  10 value 232.085041
## iter  20 value 231.385399
## iter  20 value 231.385399
## final  value 231.385399
## converged
##              Df      AIC
## + +Edad       6 450.5886
## + +IMC        9 480.7708
```

```
## + +Colesterol  6 482.4045
## <none>         3 484.2015
## # weights:  12 (6 variable)
## initial  value 277.258872
## iter  10 value 219.351589
## final  value 219.294302
## converged
##
## Step:  AIC=450.59
## Presion ~ Edad
##
## trying - Edad
## # weights:  8 (3 variable)
## initial  value 277.258872
## final  value 239.100763
## converged
## trying + Colesterol
## # weights:  16 (9 variable)
## initial  value 277.258872
## iter  10 value 218.912397
## final  value 218.133323
## converged
## trying + IMC
## # weights:  20 (12 variable)
## initial  value 277.258872
## iter  10 value 214.545138
## iter  20 value 213.412789
## final  value 213.404317
## converged
##                Df      AIC
## <none>          6 450.5886
## + +IMC         12 450.8086
## + +Colesterol   9 454.2666
## - Edad          3 484.2015
```

```
summary(Ajuste.Multinom.Step)
```

```
## Call:
## multinom(formula = Presion ~ Edad, data = Chapman.Cuali)
##
## Coefficients:
```

```
##                (Intercept)         Edad
## Normal          -0.6734662 0.0004024014
## Alta            -9.0338734 0.1643249548
## Descompensada  -1.4305665 0.0485526253
##
## Std. Errors:
##                (Intercept)       Edad
## Normal           0.8955090 0.02258743
## Alta             1.7873158 0.03364646
## Descompensada   0.6849982 0.01651098
##
## Residual Deviance: 438.5886
## AIC: 450.5886
```

El método stepwise sólo selecciona a la edad como predictora de la presión arterial.

1.2.3. Ajuste de regresión logit de respuesta nominal con observaciones en una tabla de frecuencias

Supongamos que se quiere ajustar un modelo de respuesta múltiple nominal para predecir las categorías de presión arterial a partir de las categorías del IMC, estando la información en una tabla de frecuencias del siguiente modo:

	Presión			
IMC	Óptima	Normal	Alta	Descompensada
Normal	41	16	8	46
Sobrepeso	15	12	8	44
Obesidad	0	1	1	8

donde se muestra el número de individuos de cada combinación de categorías de las variables implicadas.

Realmente se trataría de un ajuste de un modelo de regresión de respuesta múltiple multinomial simple. Para realizar el ajuste con esta información, los datos deben estar en un *Data.Frame* con el siguiente formato:

```
##           Presion       IMC Frecuencia
## 1          Optima    Normal         41
## 2          Optima Sobrepeso         15
## 3          Optima  Obesidad          0
```

```
## 4         Normal    Normal         16
## 5         Normal Sobrepeso         12
## 6         Normal  Obesidad          1
## 7           Alta    Normal          8
## 8           Alta Sobrepeso          8
## 9           Alta  Obesidad          1
## 10 Descompensada    Normal         46
## 11 Descompensada Sobrepeso         44
## 12 Descompensada  Obesidad          8
```

Supondremos que hemos llamado a este *Data.Frame* `Chapman.Tabla.Frame`

En este ejemplo se mostrarán los mismos resultados vistos en las secciones anteriores, pero en este caso sin las interpretaciones. Sólo aquellas cuestiones que difieran en algún aspecto con respecto a lo anterior se explicarán con cierto detalle.

```
Ajuste.Multinom.Tab<-multinom(Presion~IMC,
data=Chapman.Tabla.Frame,weights=Frecuencia)

## # weights:  16 (9 variable)
## initial  value 277.258872
## iter  10 value 232.085041
## iter  20 value 231.385399
## iter  20 value 231.385399
## final  value 231.385399
## converged

summary(Ajuste.Multinom.Tab)

## Call:
## multinom(formula = Presion ~ IMC, data = Chapman.Tabla.Frame,
##     weights = Frecuencia)
##
## Coefficients:
##               (Intercept) IMCSobrepeso IMCObesidad
## Normal         -0.9409597    0.7178611    7.611690
## Alta           -1.6341257    1.0055296    8.305021
## Descompensada   0.1150854    0.9610598    8.635113
##
## Std. Errors:
##               (Intercept) IMCSobrepeso IMCObesidad
```

```
## Normal          0.2947705    0.4867110    28.11029
## Alta            0.3865116    0.5840021    28.11140
## Descompensada   0.2147778    0.3681363    28.09400
##
## Residual Deviance: 462.7708
## AIC: 480.7708
```

El resultado del ajuste con la función `summary()` muestra los parámetros estimados y sus errores estándar.

Al igual que en el resto de modelos logit, la exponencial de los parámetros estimados se pueden interpretar en términos de cocientes de ventajas:

```
exp(summary(Ajuste.Multinom.Tab)$coefficients)

##               (Intercept) IMCSobrepeso IMCObesidad
## Normal          0.3902531     2.050044    2021.692
## Alta            0.1951229     2.733355    4044.129
## Descompensada   1.1219693     2.614466    5625.771
```

Las interpretaciones serán análogas a las explicadas en la sección anterior, pero sin tener en cuenta las variables explicativas cuantitativas (Edad y Colesterol).

La significación de la variable IMC se estudia a través de los test condicionales de razón de verosimilitudes:

```
anova(multinom(Presion~1,
data=Chapman.Tabla.Frame,weights = Frecuencia),multinom(Presion~IMC,
data=Chapman.Tabla.Frame,weights = Frecuencia))

## # weights:  8 (3 variable)
## initial  value 277.258872
## final  value 239.100763
## converged
## # weights:  16 (9 variable)
## initial  value 277.258872
## iter  10 value 232.085041
## iter  20 value 231.385399
## iter  20 value 231.385399
## final  value 231.385399
## converged
##   Model Resid. df Resid. Dev   Test    Df LR stat.    Pr(Chi)
```

```
## 1      1         33    478.2015           NA        NA         NA
## 2    IMC         27    462.7708 1 vs 2     6 15.43073 0.01715857
```

Obsérvese cómo el valor experimental del test es 15.43073 que para una Chi-cuadrado con 6 grados de libertad arroja un p-valor de 0.01715857 que indica que la variable IMC es significatia para la predicción de la Presión.

Los intervalos de confianza de los cocientes de ventajas:

```
exp(confint(Ajuste.Multinom.Tab))

## , , Normal
##
##                    2.5 %       97.5 %
## (Intercept)  2.189965e-01 6.954334e-01
## IMCSobrepeso 7.897215e-01 5.321723e+00
## IMCObesidad  2.388853e-21 1.710963e+27
##
## , , Alta
##
##                    2.5 %       97.5 %
## (Intercept)  9.147638e-02 4.162052e-01
## IMCSobrepeso 8.701467e-01 8.586170e+00
## IMCObesidad  4.768213e-21 3.430002e+27
##
## , , Descompensada
##
##                    2.5 %       97.5 %
## (Intercept)  7.364814e-01 1.709229e+00
## IMCSobrepeso 1.270647e+00 5.379490e+00
## IMCObesidad  6.863216e-21 4.611438e+27
```

Las probabilidades predichas y las categorías predichas:

```
predict(Ajuste.Multinom.Tab,type="probs")

##          Optima     Normal       Alta Descompensada
## 1   0.3693655246 0.14414605 0.07207167     0.4144168
## 2   0.1898712747 0.15190385 0.10126595     0.5569589
## 3   0.0001267264 0.09998353 0.10000006     0.7998897
## 4   0.3693655246 0.14414605 0.07207167     0.4144168
## 5   0.1898712747 0.15190385 0.10126595     0.5569589
```

```
## 6  0.0001267264 0.09998353 0.10000006     0.7998897
## 7  0.3693655246 0.14414605 0.07207167     0.4144168
## 8  0.1898712747 0.15190385 0.10126595     0.5569589
## 9  0.0001267264 0.09998353 0.10000006     0.7998897
## 10 0.3693655246 0.14414605 0.07207167     0.4144168
## 11 0.1898712747 0.15190385 0.10126595     0.5569589
## 12 0.0001267264 0.09998353 0.10000006     0.7998897

predict(Ajuste.Multinom.Tab,type="class")

##  [1] Descompensada Descompensada Descompensada Descompensada Descompensada
##  [6] Descompensada Descompensada Descompensada Descompensada Descompensada
## [11] Descompensada Descompensada
## Levels: Optima Normal Alta Descompensada
```

Obsérvese cómo en este caso en lugar de 200 probabilidades predichas, se muestran 12, tantas como filas tiene el *Data.Frame*. En realidad habría 200 sólo que cada una de ellas se repetiría tantas veces como muestra la columna `Frecuencia` del *Data.Frame*.

Para la tabla de clasificación habría que tener en cuenta cuántas veces se repite cada predicción:

```
table(rep(Chapman.Tabla.Frame$Presion,
Chapman.Tabla.Frame$Frecuencia),
rep(predict(Ajuste.Multinom.Tab,type="class"),
Chapman.Tabla.Frame$Frecuencia))

##
##                  Optima Normal Alta Descompensada
##   Optima              0      0    0            56
##   Normal              0      0    0            29
##   Alta                0      0    0            17
##   Descompensada       0      0    0            98
```

El estudio de la bondad del ajuste cuando la información se encuentra en tablas, es ligeramente diferente al caso de datos no agrupados.

En este caso para cada observación de la variable explicativa (Normal, Sobrepeso, Obesidad) se tiene una estimación de la probabilidad de cada categoría de la respuesta, dada por la frecuencia relativa por filas: modelo saturado.

Frecuencias absolutas:

	Presión				
IMC	Óptima	Normal	Alta	Descompensada	Total
Normal	41	16	8	46	111
Sobrepeso	15	12	8	44	79
Obesidad	0	1	1	8	10

Frecuencias relativas por filas (probabilidades predichas del modelo saturado):

	Presión				
IMC	Óptima	Normal	Alta	Descompensada	Total
Normal	0.3693694	0.1441441	0.0720721	0.4144144	1.00
Sobrepeso	0.1898734	0.1518987	0.1012658	0.556962	1.00
Obesidad	0	0.1	0.1	0.8	1.00

El test de bondad de ajuste de razón de verosimilitudes (ver página 19 de los apuntes del tema 2 para regresión logística) compara la función de verosimilitud de los datos (calculada utilizando las probabilidades estimadas a partir de los datos o frecuencias relativas) con la verosimilitud asumiendo el modelo de regresión de respuesta múltiple (calculada utilizando las probabilidades estimadas por el modelo de respuesta múltiple). Dicho de otro modo, compara la verosimilitud del modelo saturado con la del modelo de respuesta múltiple. El estadístico que se utiliza es el estadístico de Wilks de razón de verosimilitudes, que es menos 2 veces el estadístico de razón de verosimilitudes. Más concretamente, se resta la log-verosimilitud del modelo de respuesta múltiple (multiplicada por -2) y la log-verosimilitud del modelo saturado (multiplicada por -2).

Lamentablemente, R no proporciona el estadístico para este contraste, pero podemos calcularlo a partir de las salidas que proporciona R.

- La log-verosimilitud del modelo de respuesta múltiple (multiplicada por -2) es lo que R llama *deviance*
- Las probabilidades estimadas con los datos (modelo saturado) serían las frecuencias relativas por filas ($\widehat{p}_{s/q} = \frac{y_{s/q}}{n_q}$)
- El núcleo de la logverosimilitud del modelo saturado ($\sum_{q=1}^{Q} \sum_{s=1}^{S} y_{s/q} \ln \widehat{p}_{s/q}$) y por tanto la deviance del modelo saturado sería:

  $$-2 \times (41 \times ln(0{,}3693694) + 16 \times ln(0{,}1441441) + \cdots + 1 \times ln(0{,}1) + 8 \times ln(0{,}8))$$

  y el resultado es

```
## [1] 462.7683
```

- Restándo las dos deviances se tiene el estadístico de contraste

```
Estadistico<-Ajuste.Multinom.Tab$deviance-DevSaturado
Estadistico

## [1] 0.002534733
```

- El estadístico tiene distribución Chi-Cuadrado y sus grados de libertad son la diferencia entre el número de parámetros del modelo saturado (12) y el número de parámetros del modelo multinomial (9). Por ello el p-valor del contraste será:

```
pchisq(Estadistico,3,lower.tail = F)

## [1] 0.9999661
```

Este procedimiento nos lleva a aceptar el modelo de resuesta nominal como adecuado.

1.3. Ajuste de regresión logit de respuesta ordinal con R

1.3.1. Ajuste de regresión logit de respuesta ordinal con variables explicativas cuantitativas y observaciones en un *Data.Frame*

Si en el caso de ajuste de modelos de respuesta múltiple nominal, existen varias librerías y funciones que permiten el ajuste, para el caso de respuesta ordinal el número de librerías y funciones disponibles es aún más grande. Nos centraremos en este caso en una de las librerías, la librería `MASS` que se suele instalar automáticamente al instalar R, y la función `polr` pues su funcionamiento y objetos que devuelve, se parece mucho a los de la función `multinom` para el caso de respuesta nominal. La función `polr` permite el ajuste del modelo de respuesta ordinal de efectos homogéneos y ventajas proporcionales.

Para ilustrar el ajuste de un modelo de respuesta ordinal, utilizaremos los mismos ejemplos vistos para el caso de respuesta nominal, asumiendo que existe un orden en la variable respuesta, *presión arterial*, del conjunto de datos *Chapman.Cuali*, asumiendo el orden natural de las categorías de la variable, a saber, Y_1 =Optima, Y_2 =Normal, Y_3 =Alta, Y_4 =Descompensada ($Y_1 < Y_2 < Y_3 < Y_4$). Las variables explicativas cuantitativas son X_1 =Edad y X_2 =Colesterol

Para el ajuste del modelo de respuesta ordinal para la Presión a partir de las variables Edad y Colesterol, se utiliza la siguiente función de R.

```
library(MASS)
Ajuste.Ordinal.Cuanti<-polr(Presion~Edad+Colesterol,
data=Chapman.Cuali)

summary(Ajuste.Ordinal.Cuanti)

## Call:
## polr(formula = Presion ~ Edad + Colesterol, data = Chapman.Cuali)
##
## Coefficients:
##               Value Std. Error t value
## Edad       0.024826   0.012308   2.017
## Colesterol 0.003344   0.002293   1.458
##
## Intercepts:
##                     Value  Std. Error t value
## Optima|Normal      1.0357 0.6661     1.5548
## Normal|Alta        1.7202 0.6761     2.5444
## Alta|Descompensada 2.0798 0.6814     3.0520
##
## Residual Deviance: 467.9134
## AIC: 477.9134
```

La función `polr()` toma las transformaciones logit acumuladas para el ajuste. Para el ejemplo que nos ocupa, y adaptando la notación a la salida que proporciona la función, las transformaciones logit consideradas son:

$$L_s(x) = \ln\left[\frac{P\left\{Y \leq Y_s|x\right\}}{1 - P\left\{Y \leq Y_s|x\right\}}\right] = \ln\left[\frac{P\left\{Y \leq Y_s|x\right\}}{P\left\{Y > Y_s|x\right\}}\right], \ \forall s = 1, 2, 3$$

Las transformaciones logit $L_s(x)$ representan el logaritmo de la ventaja de respuesta menor o igual a Y_s frente a la respuesta mayor a Y_s. La modelización en términos de las variables explicativas será

$$L_s(X_1 = x_1, X_2 = x_2) = \beta_{s0} - \beta_1 x_1 - \beta_2 x_2, \ \forall s = 1, 2, 3.$$

De manera desarrollada:

$$\begin{aligned} L_1(X_1 = x_1, X_2 = x_2) &= \beta_{10} - \beta_1 x_1 - \beta_2 x_2 \\ L_2(X_1 = x_1, X_2 = x_2) &= \beta_{20} - \beta_1 x_1 - \beta_2 x_2 \\ L_3(X_1 = x_1, X_2 = x_2) &= \beta_{30} - \beta_1 x_1 - \beta_2 x_2 \end{aligned}$$

Obsérvese, que para cada variable explicativa se tiene un único parámetro: para $X_1 : \beta_1$ y para $X_2 : \beta_2$. También hay tres parámetros independientes que verifican que: $\beta_{10} < \beta_{20} < \beta_{30}$. Obsérvese además cómo las combinaciones lineales de las variables explicativas en esta función se definen con signo contrario a las vistas hasta este momento. Este aspecto hay que tenerlo en cuenta a efectos de interpretación de parámetros.

Los parámetros estimados junto con los elementos para el análisis de la significación de parámetros (errores estándar de estimación y valores experimentales del test de Wald) se obtienen con la matriz siguiente:

```
summary(Ajuste.Ordinal.Cuanti)$coefficients

##                          Value  Std. Error  t value
## Edad               0.024826121 0.012308185 2.017042
## Colesterol         0.003343761 0.002292831 1.458355
## Optima|Normal      1.035732148 0.666141249 1.554824
## Normal|Alta        1.720231125 0.676086539 2.544395
## Alta|Descompensada 2.079765809 0.681435704 3.052035
```

En la notación que hemos establecido los parámetros estimados son:

- $\widehat{\beta}_{10}$ =1.0357321
- $\widehat{\beta}_{20}$ =1.7202311
- $\widehat{\beta}_{30}$ =2.0797658
- $\widehat{\beta}_{1}$ =0.0248261
- $\widehat{\beta}_{2}$ =0.0033438

Obsérvese que esta salida muestra los errores estándar de estimación de los parámetros y los valores experimentales de los test de significación de los parámetros, bajo el nombre de `t value`. Sin embargo no se muestran los p-valores de esos test de significación, para lo cual podríamos utilizar la siguiente sentencia que asume distribución normal de los estimadores de los parámetros:

```
2*pnorm(abs(summary(Ajuste.Ordinal.Cuanti)$coefficients[,"t value"]),
        lower.tail=F)

##              Edad        Colesterol      Optima|Normal        Normal|Alta
##       0.043691170       0.144742769        0.119988082        0.010946726
## Alta|Descompensada
##       0.002272954
```

Al igual que en el resto de modelos logit, mediante la exponencial de los parámetros estimados se pueden obtener interpretaciones en términos de cocientes de ventajas, sin embargo en este caso se debe tener ciertas precauciones por los signos que se asumen en el modelo. En este caso el cociente de ventajas de que la respuesta tome valores menores que una cierta categoría frente a que tome valores mayores, será la exponencial del parámetro, pero cambiada de signo.

```
exp(-summary(Ajuste.Ordinal.Cuanti)$coefficients
    [c("Edad","Colesterol"),"Value"])

##       Edad Colesterol
##  0.9754795  0.9966618
```

- $\exp(-\widehat{\beta}_1)$ =0.9754795 es el cociente de ventajas de presión menor o igual a óptima frente a mayor cuando un individuo aumenta su edad en 1 año permaneciendo constante el nivel de colesterol. Equivalentemente, sería el cambio multiplicativo que se produce en la ventaja de presión menor o igual a óptima frente a mayor, cuando se aumenta un año la edad, permaneciendo constante el nivel de colesterol.

- $\exp(-\widehat{\beta}_2)$ =0.9966618 es el cociente de ventajas de presión menor o igual a óptima frente a mayor cuando un individuo aumenta su nivel de colesterol en 1 unidad permaneciendo constante la edad. Equivalentemente, sería el cambio multiplicativo que se produce en la ventaja de presión menor o igual a óptima frente a mayor, cuando se aumenta el nivel de colesterol una unidad, permaneciendo constante la edad.

- Al ser el modelo de ventajas proporcionales estos cocientes de ventajas son los mismos para presiones menores o iguales a normal o alta frente a mayores.

Sin embargo las ventajas asociadas a los términos independientes no cambian, esto es, no hay que considerar cambio de signo:

```
exp(summary(Ajuste.Ordinal.Cuanti)$coefficients
[c("Optima|Normal","Normal|Alta","Alta|Descompensada"),"Value"])

##      Optima|Normal        Normal|Alta Alta|Descompensada
##           2.817168           5.585819           8.002595
```

- $\exp(\widehat{\beta}_{10})$ =2.8171681 es la ventaja de presión menor o igual a óptima frente a mayor para individuos con Edad=0 y Colesterol=0. Esta interpretacón no tiene sentido pues en la muestra no hay individuos con Edad=0 y Colesterol=0
- $\exp(\widehat{\beta}_{20})$ =5.5858193 es la ventaja de presión menor o igual a normal frente a mayor para individuos con Edad=0 y Colesterol=0. Esta interpretacón no tiene sentido pues en la muestra no hay individuos con Edad=0 y Colesterol=0
- $\exp(\widehat{\beta}_{30})$ =8.0025946 es la ventaja de presión menor o igual a alta frente a mayor para individuos con Edad=0 y Colesterol=0. Esta interpretacón no tiene sentido pues en la muestra no hay individuos con Edad=0 y Colesterol=0.

La estimación de los parámetros se ha interpretado sin tener en cuenta la significación de estos parámetros.

También se podrían utilizar los intervalos de confianza de los parámetros para estudiar su significación:

```
confint.default(Ajuste.Ordinal.Cuanti)

##                    2.5 %      97.5 %
## Edad         0.0007025221 0.048949721
## Colesterol -0.0011501051 0.007837628
```

A partir de la estimación de parámetros, se pueden obtener las estimaciones de la probabilidades acumuladas de cada categoría de la respuesta.

Obsérvese que las probabilidades acumuladas predichas por el modelo se pueden obtener del siguiente modo para observaciones x_1 de la variable X_1 y x_2 de la variable X_2:

$$P\left\{Y \leq Y_s | x_1, x_2\right\} = \frac{\exp\left(\widehat{\beta}_{s0} + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2\right)}{1 + \exp\left(\widehat{\beta}_{s0} + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2\right)} \quad \forall s = 1, 2, 3,$$

A partir de las probabilidades acumuladas, las probabilidades de cada categoría se obtienen como:

$$\begin{aligned} P\left\{Y = Y_1 | x_1, x_2\right\} &= P\left\{Y \leq Y_1 | x_1, x_2\right\} \\ P\left\{Y = Y_s | x_1, x_2\right\} &= P\left\{Y \leq Y_s | x_1, x_2\right\} - P\left\{Y \leq Y_{s-1} | x_1, x_2\right\}, s = 2, 3 \\ P\left\{Y = Y_4 | x_1, x_2\right\} &= 1 - P\left\{Y \leq Y_3 | x_1, x_2\right\} \end{aligned}$$

Las probabilidades predichas no acumuladas de cada categoría de la respuesta, en cada valor/es de la variable/s explicativa/s las proporciona el objeto de tipo `polr()` con la sentencia `predict()` y la opción `type="probs"`. Para las 10 primeras observaciones del fichero de datos son:

```
predict(Ajuste.Ordinal.Cuanti,type="probs")[1:10,]

##         Optima    Normal       Alta Descompensada
## 1  0.2878329 0.1570338 0.08960381     0.4655295
## 2  0.3462226 0.1659785 0.08849078     0.3993081
## 3  0.2859679 0.1566487 0.08958509     0.4677983
## 4  0.3354793 0.1647690 0.08891991     0.4108318
## 5  0.1777091 0.1222595 0.08041681     0.6196145
## 6  0.2117178 0.1357694 0.08528040     0.5672323
## 7  0.2586383 0.1502509 0.08885467     0.5022561
## 8  0.1560144 0.1122025 0.07609134     0.6556918
## 9  0.2155496 0.1371273 0.08569838     0.5616247
## 10 0.2006762 0.1316740 0.08393549     0.5837144
```

Obsérvese cómo las probabilidades por filas suman la unidad como es de esperar en una distribución multinomial.

Una vez estimadas las probabilidades de cada categoría, el modelo predice, para cada valor/es de la variable/s explicativa/s la categoría de la respuesta con mayor probabilidad. Estas predicciones también las proporciona la función `polr()` con la sentencia `predict()` con la opción `type=çlass"`. Para las 10 primeras observaciones del fichero de datos son:

```
predict(Ajuste.Ordinal.Cuanti,type="class")[1:10]

##  [1] Descompensada Descompensada Descompensada Descompensada Descompensada
##  [6] Descompensada Descompensada Descompensada Descompensada Descompensada
## Levels: Optima Normal Alta Descompensada
```

Estas predicciones permiten obtener una tabla de clasificación que tiene por filas las observaciones y por columnas las predicciones:

```
table(Chapman.Cuali$Presion,
predict(Ajuste.Ordinal.Cuanti,type="class"))

##
##                 Optima Normal Alta Descompensada
##   Optima            14      0    0            42
##   Normal             7      0    0            22
##   Alta               0      0    0            17
##   Descompensada     12      0    0            86
```

Obsérvese cómo con este modelo:

- Entre los individuos con presión óptima, el modelo acierta en un 25 por ciento.
- Entre los individuos con presión normal, el modelo acierta en un 0 por ciento
- Entre los individuos con presión alta, el modelo acierta en un 0 por ciento
- Entre los individuos con presión descompensada, el modelo acierta en un 87.755102 por ciento
- Entre todos los casos, el modelo acierta en un 50 por ciento

Como en el caso de respuesta nominal, el estudio de la bondad del ajuste del modelo multinomial sólo es posible realizarlo comparando con el modelo saturado, esto es el que tiene tantos parámetros estimados como observaciones. En el caso de datos no agrupados la log-verosimilitud del modelo saturado es nula, por tanto, para estudiar la bondad del ajuste del modelo multinomial en este caso el valor experimental del test (estadístico) será la *deviance* del modelo y el p-valor el de la Chi cuadrado correspondiente:

```
Ajuste.Ordinal.Cuanti$deviance

## [1] 467.9134

pchisq(Ajuste.Ordinal.Cuanti$deviance,595,lower.tail = F)

## [1] 0.9999633
```

Hay que tener en cuenta que el modelo saturado tiene 3 parámetros libres (probabilidades) en cada combinación diferente de observaciones de las variables explicativas. En nuestro ejemplo, se tienen 200 combinaciones diferentes de observaciones de las variables explicativas (tantas como observaciones o casos) por lo que el número de parámetros libres del modelo saturado es 600. Como el modelo ajustado tiene 5 parámetros, los grados de libertad de la distribución Chi-Cuadrado del test de bondad de ajuste serán 595. Con esto, aceptamos el modelo multinomial como modelo adecuado para estos datos, frente al saturado.

1.3.2. Ajuste de regresión logit de respuesta ordinal con variables explicativas cuantitativas y cualitativas y observaciones en un *Data.Frame*

Supongamos que se quiere modelizar la presión arterial en función de la edad, el nivel de colesterol (ambas cuantitativas) y el IMC (cualitativa). En este ejemplo se mostrarán los mismos resultados vistos en la sección anterior, pero en este caso sin las interpretaciones. Sólo aquellas cuestiones que difieran en algún aspecto con respecto a lo anterior se explicarán con cierto detalle.

```
Ajuste.Ordinal.Cuanli<-polr(Presion~Edad+Colesterol+IMC,
data=Chapman.Cuali)

summary(Ajuste.Ordinal.Cuanli)$coefficients

##                          Value  Std. Error  t value
## Edad                0.017383501 0.012659910 1.373114
## Colesterol          0.003933467 0.002334219 1.685132
## IMCSobrepeso        0.629469264 0.284469245 2.212785
## IMCObesidad         1.829258543 0.809370312 2.260101
## Optima|Normal       1.187149443 0.671060867 1.769064
## Normal|Alta         1.896955880 0.682634381 2.778875
## Alta|Descompensada  2.269067770 0.688605938 3.295161
```

El resultado del ajuste con la función `summary()` muestra los parámetros estimados y sus errores estándar.

En este caso en la modelización de las transformaciones logit generalizadas se incluyen además las variables de diseño de la variable cualitativa IMC:

$$L_s(x) = \beta_{s0} - \beta_1 X_1 - \beta_2 X_2 - \tau_{s1} X_{31} - \tau_{s2} X_{32},\ \forall s = 1, 2, 3$$

donde se ha denotado por X_{31} y X_{32} a las variables de diseño de la variable IMC.

La interpretación de parámetros asociados a las categorías de IMC siguen un razonamiento similar al explicado para la Edad y el Colesterol, pero en lugar de para incrementos de la variable de una unidad, para paso de una categoría a otra de IMC. Habría que tener presente también el cambio de signo de los parámetros antes de las exponenciales.

Como ya se indicó en el caso de regresión logit binaria con variables explicativas cualitativas, la significación de la variable IMC no se puede estudiar a partir de la significación de los parámetros, sino a través de los test condicionales de razón de verosimilitudes. En presencia de las variables cuantitativas Edad y Colesterol, la significación de la variable IMC se estudiaría del siguiente modo:

```
anova(polr(Presion~Edad+Colesterol+IMC,
data=Chapman.Cuali),polr(Presion~Edad+Colesterol,
data=Chapman.Cuali))

## Likelihood ratio tests of ordinal regression models
##
## Response: Presion
##                      Model Resid. df Resid. Dev   Test    Df LR stat.
## 1        Edad + Colesterol       195   467.9134
## 2 Edad + Colesterol + IMC       193   458.1828 1 vs 2     2  9.73053
##       Pr(Chi)
## 1
## 2 0.007709785
```

Obsérvese cómo el valor experimental del test es 9.73053 que para una Chi-cuadrado con 2 grados de libertad arroja un p-valor de 0.007709785, que al 5 % de significación, indica que la variable IMC debe estar en el modelo (en presencia de la edad y el colesterol).

Las probabilidades predichas y las categorías predichas (las 10 primeras):

```
predict(Ajuste.Ordinal.Cuanli,type="probs")[1:10,]

##       Optima     Normal       Alta Descompensada
## 1  0.2303574 0.14800899 0.09057964     0.5310540
## 2  0.2699701 0.15926616 0.09253597     0.4782278
## 3  0.2222710 0.14529469 0.08989678     0.5425375
## 4  0.3848900 0.17506093 0.08869096     0.3513581
## 5  0.1490901 0.11361632 0.07806978     0.6592238
## 6  0.1845108 0.13061274 0.08518929     0.5996872
## 7  0.3208436 0.16913493 0.09227242     0.4177491
## 8  0.1233573 0.09913381 0.07087403     0.7066348
## 9  0.2816752 0.16197501 0.09272320     0.4636266
## 10 0.2606928 0.15692399 0.09226879     0.4901144

predict(Ajuste.Ordinal.Cuanli,type="class")[1:10]

##  [1] Descompensada Descompensada Descompensada Optima        Descompensada
##  [6] Descompensada Descompensada Descompensada Descompensada Descompensada
## Levels: Optima Normal Alta Descompensada
```

La tabla de clasificación:

```
table(Chapman.Cuali$Presion,
predict(Ajuste.Ordinal.Cuanli,type="class"))

##
##                  Optima Normal Alta Descompensada
##   Optima             26      0    0            30
##   Normal              5      0    0            24
##   Alta                0      0    0            17
##   Descompensada      14      0    0            84
```

La bondad del ajuste:

```
Ajuste.Ordinal.Cuanli$deviance

## [1] 458.1828

pchisq(Ajuste.Ordinal.Cuanli$deviance,593,lower.tail = F)

## [1] 0.9999888
```

Que indica que el modelo de respuesta nominal es adecuado.

Antes de terminar con este apartado vamos a mostrar el resultado de una selección stepwise de estas variables en un modelo multinomial:

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

## Start:  AIC=484.2
## Presion ~ 1
##
##              Df    AIC
## + IMC         2 476.33
## + Edad        1 478.10
## + Colesterol  1 480.04
## <none>          484.20
##
## Step:  AIC=476.33
## Presion ~ IMC
##
##              Df    AIC
## + Colesterol  1 472.08
## + Edad        1 473.12
## <none>          476.33
## - IMC         2 484.20
##
## Step:  AIC=472.08
## Presion ~ IMC + Colesterol
##
##              Df    AIC
## <none>          472.08
## + Edad        1 472.18
## - Colesterol  1 476.33
## - IMC         2 480.04

summary(Ajuste.Ordinal.Step)

## Call:
## polr(formula = Presion ~ IMC + Colesterol, data = Chapman.Cuali)
##
```

```
## Coefficients:
##                 Value Std. Error t value
## IMCSobrepeso 0.684772   0.281734   2.431
## IMCObesidad  1.981778   0.804528   2.463
## Colesterol   0.005253   0.002145   2.449
##
## Intercepts:
##                     Value  Std. Error t value
## Optima|Normal       0.8559 0.6265     1.3662
## Normal|Alta         1.5513 0.6341     2.4465
## Alta|Descompensada  1.9193 0.6391     3.0033
##
## Residual Deviance: 460.0785
## AIC: 472.0785
```

El método stepwise selecciona las variables IMC y Colesterol como predictoras de la presión arterial.

1.3.3. Ajuste de regresión logit de respuesta ordinal con observaciones en una tabla de frecuencias

Supongamos que se quiere ajustar un modelo de respuesta múltiple nominal para predecir las categorías de presión arterial a partir de las categorías del IMC, estando la información en una tabla de frecuencias del siguiente modo:

	Presión			
IMC	Óptima	Normal	Alta	Descompensada
Normal	41	16	8	46
Sobrepeso	15	12	8	44
Obesidad	0	1	1	8

donde se muestra el número de individuos de cada combinación de categorías de las variables implicadas.

Realmente se trataría de un ajuste de un modelo de regresión de respuesta múltiple multinomial simple. Para realizar el ajuste con esta información, los datos deben estar en un *Data.Frame* con el siguiente formato:

```
##         Presion       IMC Frecuencia
## 1        Optima    Normal         41
## 2        Optima Sobrepeso         15
```

```
## 3          Optima  Obesidad          0
## 4          Normal    Normal         16
## 5          Normal Sobrepeso         12
## 6          Normal  Obesidad          1
## 7            Alta    Normal          8
## 8            Alta Sobrepeso          8
## 9            Alta  Obesidad          1
## 10 Descompensada    Normal         46
## 11 Descompensada Sobrepeso         44
## 12 Descompensada  Obesidad          8
```

Supondremos que hemos llamado a este *Data.Frame* `Chapman.Tabla.Frame`

En este ejemplo se mostrarán los mismos resultados vistos en las secciones anteriores, pero en este caso sin las interpretaciones. Sólo aquellas cuestiones que difieran en algún aspecto con respecto a lo anterior se explicarán con cierto detalle.

```
Ajuste.Ordinal.Tab<-polr(Presion~IMC,
data=Chapman.Tabla.Frame,weights=Frecuencia)
summary(Ajuste.Ordinal.Tab)$coefficients

##                            Value Std. Error    t value
## IMCSobrepeso          0.70094252  0.2792344  2.5102299
## IMCObesidad           1.91051963  0.7992452  2.3904050
## Optima|Normal        -0.61161995  0.1922501 -3.1813767
## Normal|Alta           0.06221303  0.1871139  0.3324875
## Alta|Descompensada    0.41940938  0.1889788  2.2193456
```

El resultado del ajuste con la función `summary()` muestra los parámetros estimados y sus errores estándar.

Al igual que en el resto de modelos logit, la exponencial de los parámetros estimados se pueden interpretar en términos de cocientes de ventajas:

```
exp(-summary(Ajuste.Ordinal.Tab)$coefficients)

##                          Value Std. Error     t value
## IMCSobrepeso         0.4961175  0.7563626  0.08124956
## IMCObesidad          0.1480035  0.4496683  0.09159258
## Optima|Normal        1.8434152  0.8251005 24.07988182
## Normal|Alta          0.9396827  0.8293493  0.71713764
## Alta|Descompensada 0.6574350  0.8278040  0.10868021
```

Las interpretaciones serán análogas a las explicadas en la sección anterior, pero sin tener en cuenta las variables explicativas cuantitativas (Edad y Colesterol).

La significación de la variable IMC se estudia a través de los test condicionales de razón de verosimilitudes:

```
anova(polr(Presion~1,
data=Chapman.Tabla.Frame,weights = Frecuencia),polr(Presion~IMC,
data=Chapman.Tabla.Frame,weights = Frecuencia))

## Likelihood ratio tests of ordinal regression models
##
## Response: Presion
##   Model Resid. df Resid. Dev   Test    Df LR stat.   Pr(Chi)
## 1     1       197   478.2015
## 2   IMC       195   466.3348 1 vs 2     2 11.86669 0.0026496
```

Obsérvese cómo el valor experimental del test es 15.43073 que para una Chi-cuadrado con 6 grados de libertad arroja un p-valor de 0.01715857 que indica que la variable IMC es significatia para la predicción de la Presión.

Las probabilidades predichas y las categorías predichas:

```
predict(Ajuste.Ordinal.Tab,type="probs")

##        Optima     Normal       Alta Descompensada
## 1  0.35168975 0.16385849 0.08779367     0.3966581
## 2  0.21205837 0.13347543 0.08454415     0.5699221
## 3  0.07432061 0.06175123 0.04768330     0.8162449
## 4  0.35168975 0.16385849 0.08779367     0.3966581
## 5  0.21205837 0.13347543 0.08454415     0.5699221
## 6  0.07432061 0.06175123 0.04768330     0.8162449
## 7  0.35168975 0.16385849 0.08779367     0.3966581
## 8  0.21205837 0.13347543 0.08454415     0.5699221
## 9  0.07432061 0.06175123 0.04768330     0.8162449
## 10 0.35168975 0.16385849 0.08779367     0.3966581
## 11 0.21205837 0.13347543 0.08454415     0.5699221
## 12 0.07432061 0.06175123 0.04768330     0.8162449

predict(Ajuste.Ordinal.Tab,type="class")

##  [1] Descompensada Descompensada Descompensada Descompensada Descompensada
```

```
##  [6] Descompensada Descompensada Descompensada Descompensada Descompensada
## [11] Descompensada Descompensada
## Levels: Optima Normal Alta Descompensada
```

Obsérvese cómo en este caso en lugar de 200 probabilidades predichas, se muestran 12, tantas como filas tiene el *Data.Frame*. En realidad habría 200 sólo que cada una de ellas se repetiría tantas veces como muestra la columna `Frecuencia` del *Data.Frame*.

Para la tabla de clasificación habría que tener en cuenta cuántas veces se repite cada predicción:

```
table(rep(Chapman.Tabla.Frame$Presion,
Chapman.Tabla.Frame$Frecuencia),
rep(predict(Ajuste.Ordinal.Tab,type="class"),
Chapman.Tabla.Frame$Frecuencia))

##
##                 Optima Normal Alta Descompensada
##   Optima             0      0    0            56
##   Normal             0      0    0            29
##   Alta               0      0    0            17
##   Descompensada      0      0    0            98
```

El estudio de la bondad del ajuste cuando la información se encuentra en tablas, es ligeramente diferente al caso de datos no agrupados.

En este caso para cada observación de la variable explicativa (Normal, Sobrepeso, Obesidad) se tiene una estimación de la probabilidad de cada categoría de la respuesta, dada por la frecuencia relativa por filas: modelo saturado.

Frecuencias absolutas:

	Presión				
IMC	Óptima	Normal	Alta	Descompensada	Total
Normal	41	16	8	46	111
Sobrepeso	15	12	8	44	79
Obesidad	0	1	1	8	10

Frecuencias relativas por filas (probabilidades predichas del modelo saturado):

	Presión				
IMC	Óptima	Normal	Alta	Descompensada	Total
Normal	0.3693694	0.1441441	0.0720721	0.4144144	1.00
Sobrepeso	0.1898734	0.1518987	0.1012658	0.556962	1.00
Obesidad	0	0.1	0.1	0.8	1.00

El test de bondad de ajuste de razón de verosimilitudes (ver página 19 de los apuntes del tema 2 para regresión logística) compara la función de verosimilitud de los datos (calculada utilizando las probabilidades estimadas a partir de los datos o frecuencias relativas) con la verosimilitud asumiendo el modelo de regresión de respuesta múltiple (calculada utilizando las probabilidades estimadas por el modelo de respuesta múltiple). Dicho de otro modo, compara la verosimilitud del modelo saturado con la del modelo de respuesta múltiple. El estadístico que se utiliza es el estadístico de Wilks de razón de verosimilitudes, que es menos 2 veces el estadístico de razón de verosimilitudes. Más concretamente, se resta la log-verosimilitud del modelo de respuesta múltiple (multiplicada por -2) y la log-verosimilitud del modelo saturado (multiplicada por -2).

Lamentablemente, R no proporciona el estadístico para este contraste, pero podemos calcularlo a partir de las salidas que proporciona R.

- La log-verosimilitud del modelo de respuesta múltiple (multiplicada por -2) es lo que R llama *deviance*
- Las probabilidades estimadas con los datos (modelo saturado) serían las frecuencias relativas por filas ($\widehat{p}_{s/q} = \frac{y_{s/q}}{n_q}$)
- El núcleo de la logverosimilitud del modelo saturado ($\sum_{q=1}^{Q} \sum_{s=1}^{S} y_{s/q} \ln \widehat{p}_{s/q}$) y por tanto la deviance del modelo saturado sería:

  $$-2\times(41\times ln(0{,}3693694)+16\times ln(0{,}1441441)+\cdots+1\times ln(0{,}1)+8\times ln(0{,}8))$$

  y el resultado es

  ```
  ## [1] 462.7683
  ```

- Restándo las dos deviances se tiene el estadístico de contraste

  ```
  Estadistico<-Ajuste.Ordinal.Tab$deviance-DevSaturado
  Estadistico
  ```

  ```
  ## [1] 3.566564
  ```

- El estadístico tiene distribución Chi-Cuadrado y sus grados de libertad son la diferencia entre el número de parámetros del modelo saturado (12) y el número de parámetros del modelo multinomial (9). Por ello el p-valor del contraste será:

```
pchisq(Estadistico,3,lower.tail = F)

## [1] 0.3122311
```

Este procedimiento nos lleva a aceptar el modelo de resuesta ordinal como adecuado.