Modelos de regresión logística

PID_00276229

Montserrat Guillén Estany
María Teresa Alonso Alonso

Tiempo mínimo de dedicación recomendado: 4 horas

Universitat Oberta de Catalunya

Montserrat Guillén Estany **María Teresa Alonso Alonso**

La revisión de este recurso de aprendizaje UOC ha sido coordinada por la profesora: Teresa Sancho Vinuesa

Segunda edición: septiembre 2020

Av. Tibidabo, 39-43, 08035 Barcelona
Autoría: Montserrat Guillén Estany, María Teresa Alonso Alonso
Producción: FUOC

Índice

Introducción

Los **modelos de regresión** son modelos estadísticos en los que se desea conocer la relación entre una variable dependiente, que puede ser cuantitativa o cualitativa, y una o más variables explicativas independientes, o covariables, ya sean cualitativas o cuantitativas.

De entre las diferentes opciones que existen para definir un modelo de regresión destacan el modelo de regresión lineal y el modelo de regresión logística. Teniendo en cuenta el tipo de variable que deseemos estimar (variable dependiente o respuesta) aplicaremos uno u otro. Si la variable dependiente es una variable continua, el modelo de regresión más frecuentemente utilizado es la regresión lineal, y si la variable de interés es cualitativa, se utiliza la logística.

- **Variable dependiente continua**: regresión lineal.
- **Variable dependiente cualitativa**:[1] regresión logística.

(1) Especialmente útil si es dicotómica o *dummy*.

Un ejemplo sencillo de regresión lineal simple es el que relaciona el peso de una persona con su altura. En este caso se considera que el peso (variable dependiente cuantitativa, variable respuesta, explicada o efecto) es consecuencia de la altura (variable independiente, variable tratamiento, explicativa o causa), aparte de otros muchos factores. En un modelo de regresión lineal simple, solo hay una variable dependiente y una independiente. Se establece que la relación es una recta (llamada *recta de regresión*) y que a partir de un conjunto de datos la estimación de la recta permite efectuar predicciones. La predicción del peso de una persona en función de su altura se obtiene mediante la recta estimada y fijando que dicho peso corresponde al valor que toma la recta en el punto concreto de la altura. En este ejemplo, queda patente que la predicción es un «valor esperado» que se obtiene del ajuste a un modelo muy simple que han proporcionado unos datos históricos.

El paso de la regresión lineal simple a la regresión lineal múltiple surge de la necesidad de incorporar más factores en la explicación de la variable dependiente. En el ejemplo de peso y altura se pueden tener en cuenta características como la cantidad diaria promedio de calorías diarias ingeridas, el sexo o si se realiza algún tipo de actividad física habitualmente. Por lo tanto, con un modelo de regresión lineal múltiple, al considerarse más factores, se puede realizar una predicción más precisa que en un modelo simple, si se conoce la información sobre dichas características. En el modelo múltiple hay una variable dependiente y varias independientes.

El término **error** recoge todos los factores que influyen en el valor observado de la variable dependiente no recogidos por la variables explicativas y se supone que tiene un promedio nulo (los errores a veces son positivos y otras veces negativos). En la inferencia sobre los modelos de regresión lineal, se supone que el término de error sigue un comportamiento normal. La predicción que proporciona el modelo de regresión lineal es un valor esperado de la respuesta, en función de los valores de las variables explicativas.

Una de las restricciones que tienen los modelos de regresión lineal simple o múltiple es precisamente la linealidad. Dicha hipótesis puede ser excesivamente restrictiva cuando la realidad no lo es. Para ello, los modelos no lineales establecen relaciones causa-efecto más complejas, que no solo suponen un reto para la elección de la forma de la relación (parabólica, exponencial, polinomial, etc.) sino que a su vez complican la estimación del modelo.

La segunda de las restricciones importantes de los modelos de regresión lineal clásicos es considerar que la variable dependiente tiene un comportamiento continuo; es decir, toma valores con decimales. Este es el caso del peso, que podría llegar a medirse con mucha precisión. Sin embargo, en la práctica existen numerosos fenómenos que no pueden medirse de manera tan específica, ya sea por su propia naturaleza o porque ya han sido registrados como variables no continuas.

En el texto siguiente se introduce el modelo de regresión logística, que suple las limitaciones del modelo de regresión lineal comentadas anteriormente.

Así pues, el **modelo de regresión logística** es una técnica estadística, por medio de la cual se analizan las relaciones de asociación entre una variable dependiente cualitativa o politómica (admite varias categorías de respuesta) y una o varias variables independientes (regresores o predictores), que pueden ser cuantitativas o categóricas. Este tipo de regresión es especialmente interesante si la variable que queremos estimar es dicotómica o *dummy*.

Un ejemplo de una variable no continua es tomar una decisión entre conceder un crédito a un cliente o no. En este caso tenemos una variable dicotómica, es decir, una variable que únicamente puede tomar dos valores («Sí conceder el crédito» y «No conceder el crédito»).

Como variables explicativas podemos tener en cuenta el nivel salarial medido de manera discreta, con valores 1, 2, 3 o 4, estado civil, sexo o número de hijos, entre otras.

Tabla 1. Datos de solicitud de préstamos de una entidad bancaria

ID	Edad	Sexo	Estado civil	Número de hijos	Nivel salarial	Crédito solicitado	Préstamo	Motivo
56	46	H	C	2	3	20.000	65.000	Vehículo
76	34	M	S	0	2	5.000	0	Estudios
96	52	H	C	2	4	350.000	0	Reformas

Fuente: ejemplo del módulo «Introducción a la estadística».

Estos modelos pueden acercarse más a la realidad de muchos fenómenos, ya que la relación entre las variables se asemeja más a una curva que a una recta.

Tal como ya ocurría en los modelos de regresión lineal clásicos, las variables explicativas pueden ser de cualquier tipo, es decir, continuas, cualitativas dicotómicas, politómicas o discretas. Sin embargo, será la naturaleza de la variable dependiente la que determine qué modelo predictivo es necesario utilizar.

1. Modelos lineales generalizados (MLG)

Los modelos lineales generalizados fueron inicialmente formulados en 1972 y 1974 por dos profesores ingleses, John Nelder y Robert Wedderburn, respectivamente. Sin embargo, posteriormente el impulsor fue Peter McCullagh, quien en 1983 publicó junto con John Nelder el libro *Generalized Linear Models*.

La posibilidad de utilizar una creciente potencia computacional en una década en la que se popularizó el uso de ordenadores personales permitió que los nuevos modelos pudieran estar al alcance de investigadores de todos los ámbitos, desde la medicina y la biología a las ciencias sociales, la economía y el marketing. Esto permitió una rápida expansión de dichos modelos.

Los **modelos lineales generalizados** logran unificar un gran número de modelos estadísticos predictivos, incluyendo la regresión lineal, la regresión logística (cuando la respuesta es dicotómica) y la regresión de Poisson (cuando la respuesta es un valor que cuenta el número de veces que ocurre un fenómeno).

Los modelos lineales generalizados tienen tres componentes:

1) El predictor lineal.

2) El comportamiento aleatorio de la variable dependiente.

3) Una función que da una correspondencia entre el predictor lineal y el valor esperado de la variable dependiente, que se denomina *link* o *función de ligadura*.

Solo puede considerarse que un modelo lineal generalizado está bien definido si:

- El predictor lineal es una combinación lineal entre parámetros y variables explicativas.
- El comportamiento estocástico de la variable dependiente se encuentra dentro de un conjunto concreto de distribuciones denominado *la familia exponencial*.
- La función de ligadura tiene ciertas propiedades (como ser continua, monótona y derivable).

Tabla 2. Ejemplos de MLG

Variable dependiente	Variables explicativas	Naturaleza de la variable dependiente	Distribución	Modelos lineales generalizados posibles
El consumidor decide comprar o no comprar un producto	Sexo, edad, nivel de estudios, zona de residencia, etc.	Dicotómica (toma dos valores, sí o no)	Bernoulli	Modelo de regresión logística, modelo *probit*, etc.
Otorgar un crédito o no	Edad, nivel salarial, estado civil, número de hijos	Dicotómica (toma dos valores, sí o no)	Bernoulli	Modelo de regresión logística, modelo *probit*, etc.
Número de días que se pernocta en una ciudad durante un viaje vacacional	Edad, nacionalidad, nivel socioeconómico	Discreta (los valores posibles son: 0, 1, 2, 3, etc.)	Poisson	Modelo de Poisson

Los impulsores de los modelos lineales generalizados propusieron un método de mínimos cuadrados reponderado iterativo para la estimación de máxima verosimilitud de los parámetros del modelo. De este modo se obtienen las estimaciones y sus errores para la obtención de los respectivos intervalos de confianza con un número pequeño de iteraciones, y aunque existen otras posibles vías de estimación, esta sigue siendo la más ampliamente utilizada.

Lectura recomendada

R. W. Wedderburn (1974). «Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method». *Biometrika* (vol. 61, n.º 3, págs. 439-447). Oxford: Oxford University Press.

Ejemplo de modelo lineal generalizado

El fabricante de perfumes Mi aroma favorito desea conocer cómo son los consumidores que tienen mayor propensión a comprar una fragancia que va a lanzar próximamente.

Para ello, elige un conjunto de 150 personas a las que pregunta si querrían comprarla o no.

La respuesta es dicotómica. Los factores que influyen en dicha decisión y que van a ser considerados son edad, sexo, situación laboral e ingresos netos mensuales. Veamos en primer lugar que, si bien la variable dependiente es dicotómica, las variables explicativas tienen naturaleza distinta. La edad es una variable cuantitativa discreta, el sexo es cualitativa (por lo tanto, se codifica por ejemplo tomando el valor 1 si es una mujer y 0 si es un hombre), la situación laboral considera tres posibilidades: trabajador por cuenta propia o ajena, parado u otros (jubilado, estudiante, etc.), y finalmente la variable ingresos netos mensuales se considera continua. Para la situación laboral se establece ser trabajador por cuenta propia o ajena como la categoría de referencia y, por lo tanto, se definen dos variables dicotómicas: estar parado (SL1) toma el valor 1 y 0 en caso contrario, y encontrarse en cualquier otra situación excepto parado o trabajando (SL2), que se codifica como 1 y se da un 0 si es un trabajador por cuenta propia o ajena o se está en paro.

Modelizaremos este problema mediante la definición de tres componentes:

1) El componente denominado ***predictor lineal*** de este modelo es:

$$b_0 + b_1 * \text{edad}_i + b_2 * \text{mujer}_i + b_3 * \text{SL1}_i + b_4 * \text{SL2}_i + b_5 * \text{ingresos}_i$$

Es importante destacar que:

- En este ejemplo tenemos seis parámetros (los parámetros b_i), que son los que vamos a estimar.
- Cada participante en este estudio de mercado tiene unas características propias de edad, sexo, situación laboral e ingresos, por ello se utiliza el subíndice «*i*», que indica que se utiliza el valor observado de dicha variable para la *i*-ésima persona preguntada. Suele ponerse al lado del modelo la indicación $i = 1{:}150$, lo que quiere decir que existe un predictor lineal diferente para cada uno de los participantes en el estudio. En general, se habla de *N* participantes.
- Una vez que se hayan estimado los parámetros, se podrá calcular el predictor lineal para cualquier tipo de consumidor, si se conoce su edad, sexo, situación laboral e ingresos, aunque no haya ninguno exactamente como él en el estudio inicial.

2) El **componente aleatorio** en este caso es una variable que toma dos posibles valores «Sí compraría» o «No compraría». Se suele indicar con la letra *Y* por ser variable dependiente y el subíndice «*i*» para indicar que se refiere al *i*-ésimo participante. Se puede codificar con un 1 en el caso afirmativo y 0 en caso negativo.

De este modo, para cada participante tenemos una variable aleatoria Bernouilli:

$$Y_i = \begin{cases} 0 & \text{Con probabilidad } (1-p_i) \\ 1 & \text{Con probabilidad } p_i \end{cases}$$

Respecto a este componente, hay que destacar que la probabilidad de comprar o no es única para cada persona *i* particular, de ahí que se utilice una probabilidad con el subíndice «*i*».

Además, por tratarse de una variable dicotómica, el valor esperado de la variable corresponde exactamente a la probabilidad de comprar. Es decir, la esperanza matemática de comprar para cada participante se calcula como:

$$E(Y_i) = 0 * (1 - p_i) + 1 * p_i = p_i$$

En definitiva, el valor que va a predecirse es la probabilidad de que la variable dependiente tome el valor 1, es decir:

$$E(Y_i) = 0 * (1 - p_i) + 1 * p_i = P(Y_i = 1)$$

3) El tercer componente es una **función de ligadura** que hace corresponder a la probabilidad de comprar de cada individuo un valor del predictor lineal. Esta función cumple las condiciones exigibles en el modelo lineal generalizado. Es decir, debe ser una función monótona (creciente o decreciente), continua y diferenciable. La elección de la función de ligadura determinará finalmente el tipo de modelo lineal generalizado que se haya establecido.

$$E(Y_i) = p_i = f(b_0 + b_1 * \text{edad}_i + b_2 * \text{mujer}_i + b_3 * \text{SL1}_i + b_4 * \text{SL2}_i + b_5 * \text{ingresos}_i)$$

En el caso de un **modelo de regresión logística** la elección sería la siguiente:

$$P(Y_i = 1) = \frac{\exp(b_0 + b_1 * \text{edad}_i + b_2 * \text{mujer}_i + b_3 * \text{SL1}_i + b_4 * \text{SL2}_i + b_5 * \text{ingresos}_i}{1 + \exp(b_0 + b_1 * \text{edad}_i + b_2 * \text{mujer}_i + b_3 * \text{SL1}_i + b_4 * \text{SL2}_i + b_5 * \text{ingresos}_i}$$

Todos los modelos lineales generalizados se definen mediante los tres componentes considerados y, por lo tanto, se determinan fijando:

- ¿Cuáles son las variables que participan en la determinación del predictor lineal?
- ¿Qué distribución estadística se utiliza para la variable dependiente?
- ¿Cuál es la función de ligadura?

2. Estimación por máxima verosimilitud

Para la estimación de los coeficientes del modelo de regresión logística y de sus errores estándar se recurre al cálculo de estimaciones de máxima verosimilitud, es decir, estimaciones que maximicen la probabilidad de obtener los valores de la variable dependiente *Y* proporcionados por los datos de la muestra.

La **función de verosimilitud** se define a partir de los datos observados, del predictor lineal y de la función de ligadura, así como del comportamiento estocástico de la variable dependiente.

Dicha función se basa en la independencia de las observaciones, por lo tanto, supone que las respuestas observadas no tienen factores que las afecten y que puedan inducir dependencia entre estas.

Por ejemplo, en el caso del ejemplo de la perfumería anterior no sería adecuado que dos personas opinaran a la vez sobre la fragancia porque la respuesta de una podría influir en la de su compañera.

Intuitivamente, la función de verosimilitud es aquella que asocia a cada vector de parámetros posibles la probabilidad de observar los datos disponibles, si dichos parámetros fueran ciertos. Por lo tanto, en el ejemplo que hemos considerado, se tendría una función de seis variables, cada una de las cuales tiene un dominio determinado en la recta real y su recorrido será el intervalo [0, 1], dado que el resultado es una probabilidad.

El procedimiento de estimación de un modelo lineal generalizado consiste en encontrar los parámetros que hacen mayor la probabilidad final. El resultado es, pues, un único valor estimado para cada parámetro al que se asocia un error estándar.

La idea que subyace en la maximización de la verosimilitud puede explicarse metafóricamente como la de un explorador que busca las coordenadas de la cima de una montaña. En la cumbre está el máximo y debe encontrar aquel punto de las coordenadas del mapa donde este se encuentra. Para ello, inicia un recorrido a partir de un lugar inicial y va ascendiendo por la montaña hasta hallar un lugar en el que ya no es posible subir más arriba. En ese momento, las coordenadas dan el punto exacto del máximo.

El procedimiento iterativo que propusieron Nelder y Wedderburn se basa en este principio y precisamente la forma de definir los modelos lineales generalizados garantiza que se puede encontrar dicho máximo con pocas iteraciones.

3. Modelos de regresión logística binaria

Un caso particular de los modelos lineales generalizados son los modelos de regresión logística, donde la variable dependiente es de naturaleza dicotómica y sigue una distribución binomial.

Se dice que una variable aleatoria es **binomial** cuando solo tiene dos posibles resultados: «éxito» y «fracaso», y la probabilidad de cada uno de ellos es constante en una serie de repeticiones.

Se caracteriza por la probabilidad de éxito, representada por p, y por la probabilidad de fracaso, que se representa por q; ambas probabilidades están relacionadas por la relación $p + q = 1$.

En ocasiones, se usa el cociente p/q, denominado *odds*, que indica cuánto más probable es el éxito que el fracaso.

Los **modelos de regresión logística binaria** son modelos de regresión que permiten estudiar si una variable binomial depende o no de otra u otras variables (no necesariamente binomiales).

Es frecuente encontrar situaciones donde las variables objeto de estudio son de naturaleza dicotómica o binomial; por ejemplo, en el estudio de nuevos fármacos para hallar la probabilidad de que dichos medicamentos sean efectivos en pacientes con determinadas características; en economía para predecir si una acción incrementará o reducirá su valor; en psicología para conocer si la respuesta a un estímulo será positiva o negativa; en salud pública, estado de salud (bueno o malo), etc.

Ejemplo de modelo de regresión logística binaria

Supongamos que queremos estudiar el efecto que una serie de variables independientes tienen sobre la percepción de nuestro estado de salud, como por ejemplo, edad, sexo, IMC (índice de masa corporal), consumo de alcohol y tabaco.

Se debe construir un modelo que describa la relación entre la variable dependiente dicotómica (buena o mala salud) y una o varias variables independientes explicativas (también denominadas covariables, regresores o predictores), sean cuantitativas o categóricas.

Podemos definir la variable dependiente dicotómica Y como:

$$Y_i = \begin{cases} 0 & \text{Percepción de No buena salud} \\ 1 & \text{Percepción de Sí buena salud} \end{cases}$$

donde la variable toma el valor 1 con probabilidad π y el valor 0 con probabilidad $(1 - \pi)$.

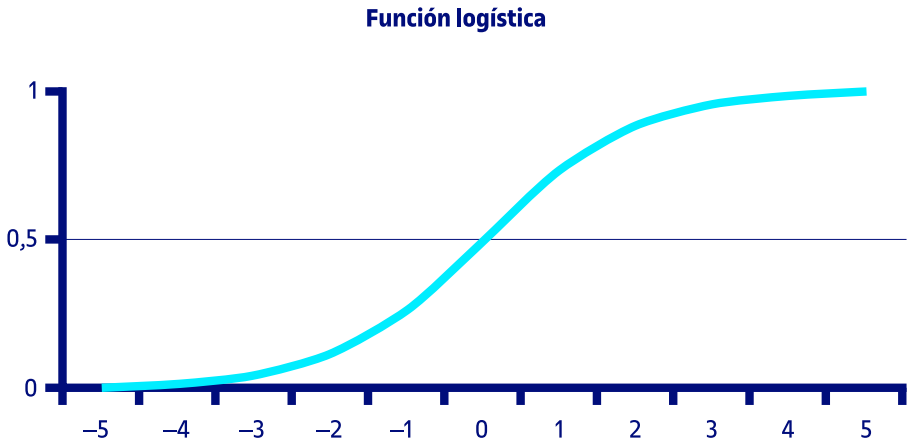
Objetivos del modelo

- Determinar la existencia o ausencia de relación entre una o más variables independientes y la variable dependiente, percepción de salud.
- Medir la magnitud de dicha relación.
- Estimar o predecir la probabilidad de que el estado de salud sea bueno o malo, en función de los valores que adopten las variables independientes.

Al clasificar el valor de la variable dependiente como 0 y 1, para determinar la relación entre ese suceso (percepción del estado de salud) y una o más variables independientes, se podría caer en el error de utilizar un modelo de regresión lineal y estimar los coeficientes por el procedimiento de mínimos cuadrados. Sin embargo, aunque esto es posible matemáticamente, los resultados obtenidos serían absurdos. Efectivamente, el valor de la función obtenida para diferentes valores de las variables independientes nos daría resultados, en general, diferentes de 0 y 1, ya que esa restricción no se impone en la regresión lineal, en la que la respuesta puede tomar cualquier valor.

Como se puede observar en la figura 1, la relación entre la variable dependiente dicotómica (0 y 1) y la variable independiente no está definida por una recta, sino que sigue una forma sigmoidea, propia de la función logística.

Figura 1. Función logística



Ecuación del modelo

El modelo de regresión logística binaria determina la probabilidad de que la variable aleatoria Y tome el valor 1 (éxito):

$$P(Y=1/X)=\frac{e^{b0}+\sum_{i=1}^{n}(b_i x_i)}{1+e^{b0}+\sum_{i=1}^{n}(b_i x_i)}$$

siendo $P(Y = 1/X)$ probabilidad de que Y tome el valor 1, en presencia de las variables independientes X.

Los parámetros de esta ecuación son:

- b_0: constante del modelo o término independiente.
- n: número de variables independientes o covariables.
- b_i: coeficientes de las covariables.

Si dividimos la expresión anterior por su complementario, es decir, si se construye su *odds*, se obtiene la expresión:

$$\frac{P(Y=1/X)}{1-P(Y=1/X)} = e^{b_0+\sum_{i=1}^{n}(b_i x_i)}$$

En nuestro ejemplo la probabilidad de percibir «Sí buena salud», entre la probabilidad de percibir «No buena salud».

Si ahora realizamos su transformación logarítmica con el logaritmo neperiano, obtenemos una ecuación lineal que es más fácil de manejar y comprender.

$$\ln\left(\frac{P(Y=1/X)}{1-P(Y=1/X)}\right) = b_0+\sum_{i=1}^{n}(b_i x_i)$$

A la izquierda de la igualdad está el *logit*, es decir, el logaritmo neperiano de la *odss* de la variable dependiente. El término a la derecha de la igualdad es la expresión de una recta, idéntica a la del modelo general de regresión lineal.

Como se comentó en el apartado anterior, para la estimación de los coeficientes del modelo y de sus errores estándar se recurre al cálculo de estimaciones de máxima verosimilitud, es decir, estimaciones que maximicen la probabilidad de obtener los valores de la variable dependiente Y, proporcionados por los datos de la muestra.

Estas estimaciones no son de cálculo directo, tal y como ocurre en el método de mínimos cuadrados. Para el cálculo de estimaciones de máxima verosimilitud se recurre a métodos iterativos, como el método de Newton-Raphson para el cálculo de ceros de funciones.

Seguimos con nuestro ejemplo. Supongamos que queremos estudiar el efecto del sexo y consumo de alcohol en nuestra percepción de salud.

Como variables independientes tenemos:

- *Sexo*: toma los valores 0 si es hombre y 1 si es mujer.
- *Bebedor*: toma los valores 0 (Poco/Nada), 1 (Ocasionalmente) y 2 (Frecuentemente).

Como variable dependiente tenemos: percepción de estado de salud.

$$Y_i = \begin{cases} 0 & \text{Percepción de No buena salud} \\ 1 & \text{Percepción de Sí buena salud} \end{cases}$$

En el programa estadístico R la función que se usa para este tipo de modelos es la función «glm()».

Los argumentos de esta función son dos: fórmula y familia (binomial, Poisson, etc.).

```
bebedor_Rel=relevel(data_salud$bebedor, ref = '0')

model.logist1=glm(formula=salud~sexo+factor(bebedor_Rel),family=binomial(link=logit))
summary(model.logist1)
##
## Call:
## glm(formula = salud ~ sexo + factor(bebedor_Rel), family = binomial(link = logit))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9851   0.5480   0.5629   0.7040   0.7814
##

## Coefficients:
##        0              Estimate Std. Error z value Pr(>|z|)
## (Intercept)            1.26858    0.06027  21.049  < 2e-16 ***
## sexo2                 -0.23868    0.06230  -3.831 0.000128 ***
## factor(bebedor_Rel)1  0.55151    0.06288   8.771  < 2e-16 ***
## factor(bebedor_Rel)2  0.49377    0.15984   3.089 0.002007 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7177.9  on 7356  degrees of freedom
## Residual deviance: 7059.2  on 7353  degrees of freedom
## AIC: 7067.2
##
## Number of Fisher Scoring iterations: 4
```

Donde se muestra la tabla con la estimación de la constante y los coeficientes del modelo (***Estimate***), sus errores estándar (***Std. Error***), el valor del estadístico de Wald (***z value***) y el *p-valor* asociado **Pr(>|z|)**), la ***null deviance***, ***residual deviance*** y **AIC**. Todos ellos se comentarán a continuación.

Por otro lado, se ha explicado más arriba que los parámetros del modelo se estiman maximizando la función de verosimilitud, mediante el método de Newton-Raphson. Las iteraciones que aparecen (***Number of Fisher Scoring iterations***) se deben a que es necesario resolver las ecuaciones de forma iterativa. En este caso han sido necesarias cuatro iteraciones para estimar los parámetros.

Según los resultados obtenidos, se muestran los valores estimados de los coeficientes del modelo y se construye el *logit*:

$$\ln\left(\frac{P(Y=1/X)}{1-P(Y=1/X)}\right) = 1{,}27 - 0{,}238 * \text{sexo} + 0{,}55 * \text{bebe } 1 + 0{,}49 * \text{bebe } 2$$

Se puede observar que tanto la variable *sexo* como la variable *bebedor* son significativas porque **Pr(>|z|) < 0,05**. Esta información la proporciona el estadístico de Wald.

Por otro lado están los errores estándar asociados con los coeficientes (***Std. Error***), y con ellos se construyen los intervalos de confianza (IC).

Los IC nos orientan sobre los posibles valores que puede tomar el verdadero valor del parámetro. Cuanto más ancho sea el intervalo, más ineficiente es la estimación del parámetro.

Para calcular los IC en el programa R, se utiliza la función `confint`:

```
confint(model.logist1)
## Waiting for profiling to be done...
##                            2.5 %     97.5 %
```

```
## (Intercept)           1.1512601  1.3875492
## sexo2                -0.3609674 -0.1167008
## factor(bebedor_Rel)1  0.4283962  0.6749056
## factor(bebedor_Rel)2  0.1891390  0.8169365
```

Para poder interpretar los resultados obtenidos anteriormente, primero hemos de definir los *odds* y *odds-ratio*.

3.1. Interpretación de los *odds-ratios* y los coeficientes del modelo

3.1.1. *Odds-ratio*

Los ***odds*** es la razón de la probabilidad de ocurrencia de un suceso entre la probabilidad de su no ocurrencia.

En esta expresión, el modelo está expresado en términos del *log-odds*:

$$\ln\left(\frac{P(Y=1/X)}{1-P(Y=1/X)}\right) = 1{,}27 - 0{,}238 * \text{sexo} + 0{,}55 * \text{bebe } 1 + 0{,}49 * \text{bebe } 2$$

Si se escribe en términos de *odds*, se tiene:

$$\frac{P(Y=1/X)}{1-P(Y=1/X)} = \frac{e^{b_0} + \sum_{i=1}^{n}(b_i x_i)}{1 + e^{b_0} + \sum_{i=1}^{n}(b_i x_i)}$$

Se calculan los distintos valores de las probabilidades para las cuatro combinaciones entre la variable dependiente Y con la independiente X:

$$\frac{P(Y=1/X=1)}{1-P(Y=1/X=1)} = \frac{e^{b_0+b_1}}{1+e^{b_0+b_1}}$$

$$\frac{P(Y=1/X=0)}{1-P(Y=1/X=0)} = \frac{e^{b_0}}{1+e^{b_0}}$$

$$\frac{P(Y=0/X=1)}{1-P(Y=0/X=1)} = \frac{1}{1+e^{b_0+b_1}}$$

$$\frac{P(Y=0/X=0)}{1-P(Y=0/X=0)} = \frac{1}{1+e^{b_0}}$$

Los *odds-ratio* (OR) se calculan como la razón entre los *odds*, donde la variable respuesta Y está presente entre los individuos, es decir, toma el valor $Y = 1$, y la variable independiente X puede estar presente o no, es decir, tomar los valores $X = 1$ y $X = 0$.

$$\mathrm{OR} = \frac{\frac{P(Y=1/X=1)}{1-P(Y=1/X=1)}}{\frac{P(Y=1/X=0)}{1-P(Y=1/X=0)}} = e^{b1}$$

- Un OR = 1 implica que no existe asociación entre la variable respuesta y la covariable.

- Un OR inferior a la unidad se interpreta como un factor de protección, es decir, el suceso es menos probable en presencia de dicha covariable.

- Un OR mayor a la unidad se interpreta como un factor de riesgo, es decir, el suceso es más probable en presencia de dicha covariable.

Supongamos que queremos calcular el OR para la covariable (*sexo*), a partir de los datos recogidos en la tabla de contingencia siguiente:

```
table(salud,sexo)
##                   sexo
## salud            H    M
##  Si buena salud 3072  2879
##  No buena salud 594   812
```

En este caso tenemos dos variables: *sexo* y *percepción de salud*.

- *Salud*: variable dependiente percepción de salud con valores «Sí buena salud» y «No buena salud».
- *Sexo*: es la covariable con valores H (hombre) y M (mujer).

Se observa que, entre los hombres, hay 594 que creen que tienen mala salud frente a 3.072 que piensan que su salud es buena. Respecto a las mujeres, hay 812 frente a 2.879. Si lo pasamos a porcentajes, se tendría para el caso de los hombres un 83,8 % que percibe buena salud frente a un 16,2 % que no; y en las mujeres un 78 % frente a un 22 %.

Si calculamos el *odds* de percibir mala salud respecto a percibir buena salud, se tiene:

- *Odds* (hombres) = 16,2/83,8 = 0,193. Por lo tanto, de cada 10 hombres que perciben buena salud, 2 no la perciben.

- *Odds* (mujeres) = 22/78 = 0,282. Es decir, de cada 10 mujeres que perciben buena salud, 3 no la perciben.

Así pues, la percepción sobre su salud del género femenino es peor que la del masculino. ¿Cuánto peor?

Si calculamos el OR a partir de la fórmula dada anteriormente se tiene:

$$\mathrm{OR} = \frac{0,282}{0,193} = 1,46$$

Expresado en términos de probabilidad significaría que la probabilidad de encontrar a una mujer que sienta que tiene mala salud, sobre una que no, es de 1,46 veces respecto al caso de los hombres. Por lo tanto, la probabilidad es un 46 % mayor con respecto a la de los hombres.

Nótese que al ser el valor de la OR mayor a 1, significaría que el hecho de ser mujer es un factor de riesgo para la percepción que tenemos sobre nuestra salud.

Es muy importante en la interpretación de los resultados que no estamos hablando de que una persona esté enferma o no, sino de cómo ella se siente.

3.1.2. Coeficientes del modelo

A la hora de interpretar los **coeficientes del modelo** hay que tener en cuenta dos cosas: la relación funcional entre la variable dependiente y las independientes, y la unidad de cambio para las variables independientes. Además, la interpretación de los coeficientes dependerá también de qué tipo de variables independientes tengamos: dicotómicas, politómicas o continuas.

Un primer paso útil es analizar el signo de los parámetros y comprobar si los signos estimados son aquellos que la intuición previa o la teoría indicaban.

Un parámetro positivo significa que un aumento en la variable que está asociada a este parámetro implica un aumento en la probabilidad de la respuesta del suceso analizado. Por el contrario, si un parámetro es negativo, cuando el factor predictivo (variable independiente) aumenta la probabilidad del suceso modelado disminuye.

Los modelos de regresión logística utilizados en la práctica siempre deben contener un término constante. El valor de la constante no es directamente interpretable. Se utiliza como un valor promedio y corresponde al logaritmo natural (o logaritmo neperiano) de la probabilidad del suceso cuando todos los regresores son iguales a cero.

1) Caso 1: variable independiente dicotómica

Supongamos que con los datos sobre la percepción del estado de salud queremos ver la relación entre esta variable y el sexo.

Como primer paso se introducirá en el modelo solo la variable explicativa *sexo*. Esto es lo que se denomina *el análisis crudo*, es decir, sin ajustar por ninguna otra variable explicativa más. Posteriormente, se estimará el OR crudo.

```
model.logist2=glm(formula=salud~sexo,family=binomial(link=logit))
summary(model.logist2)


##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.64320    0.04482  36.661  < 2e-16 ***
## sexo2       -0.37751    0.05990  -6.302 2.93e-10 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En los resultados obtenidos se observa que el signo del coeficiente asociado al *sexo* es negativo, lo que significa que la percepción de salud de las mujeres es peor que la de los hombres.

Si queremos conocer la magnitud de dicha relación, debemos calcular el OR.

Utilizando el programa R, se ejecutará el comando `exp()`, con el cual se calculará la exponencial de los coeficientes obtenidos anteriormente.

```
exp(coefficients(model.logist2))
## (Intercept)       sexo2
##   5.1717172   0.6855685
```

Junto al estimador del OR, es importante utilizar los intervalos de confianza (IC) para obtener información adicional acerca del parámetro. Los IC para los coeficientes del modelo se calculan asumiendo que la distribución de los parámetros estimados es aproximadamente normal. De este modo, para calcular los IC de los OR basta con aplicar la función exponencial `exp()` a los IC de los coeficientes estimados en el modelo.

```
exp(confint(model.logist2))
## Waiting for profiling to be done...
##                 2.5 %   97.5 %
## (Intercept) 4.7408181 5.651618
## sexo2       0.6094711 0.770800
```

Se tiene un OR para la variable *sexo* de 0,68, con lo que la ocurrencia de percibir buena salud en las mujeres es 0,68 veces menor, en relación con los hombres.

Además, como el IC para el OR de *sexo* es (0,61; 0,77), se puede decir que la buena salud entre las mujeres en la población de estudio es entre 0,61 y 0,77 veces menos probable que en los hombres.

2) Caso 2: variable independiente politómica

En este caso la variable independiente tiene más de dos categorías o, lo que es lo mismo, toma más de dos valores.

En nuestro ejemplo teníamos tipo de bebedor, que toma los valores 0 (Poco/Nada), 1 (Ocasionalmente) y 2 (Frecuentemente).

R crea tres variables dicotómicas que indican si la persona es un tipo de bebedor u otro (variables *dummy* o ficticias).

Al ajustar el modelo, R está suponiendo que el valor de la primera variable es 0 y por lo tanto está comparando los otros dos grupos con poco/nada bebedor.

Si se efectúa el análisis crudo, se obtiene:

```
model.logist3=glm(formula=salud~factor(bebedor_Rel),family=binomial(link=logit))
summary(model.logist3)
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            1.10804    0.04243  26.117  < 2e-16 ***
## factor(bebedor_Rel)1  0.61136    0.06094  10.031  < 2e-16 ***
## factor(bebedor_Rel)2  0.57691    0.15825   3.645 0.000267 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
exp(coefficients(model.logist3))
##          (Intercept) factor(bebedor_Rel)1 factor(bebedor_Rel)2
##             3.028417             1.842933             1.780520
```

El *odds-ratio* estimado para `bebedor_Rel 1` es de 1,84, por lo que podemos concluir que la probabilidad de percibir buena salud para personas que beben moderadamente, comparado con los no bebedores, es 1,84 veces mayor.

De manera análoga, la probabilidad de que las personas que consumen mucho alcohol crean que tienen buena salud, respecto a los no bebedores, es 1,78 veces mayor.

En consecuencia, la ocurrencia de percibir buena salud en las personas que consumen alcohol es mayor con relación a las que no consumen. Se puede decir que las personas que consumen alcohol son más optimistas, respecto a su salud, aunque en realidad sea lo contrario.

3) Caso 3: variable independiente continua

Cuando se utiliza una variable predictora continua, la interpretación del parámetro estimado depende de las unidades en las que se mide.

$$\ln\left(\frac{P(Y=1/X)}{1-P(Y=1/X)}\right)=b_0+\sum_{i=1}^{n}b_i x_i$$

El coeficiente b_i representa el cambio en el ln (*odds*) cuando la variable X cambia en una unidad, o equivalentemente, e^{b_i} indica la variación en el *odds* cuando la variable X cambia en una unidad.

Si en vez de variar en una unidad lo hace en c unidades, entonces el cambio vendrá dado por e^{c*b_i}.

Vamos a establecer la relación entre el índice de masa corporal y la percepción de salud.

```
model.logist5=glm(formula=salud~data_salud$imc,family=binomial(link=logit)
summary(model.logist5)

## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     4.120383   0.195110   21.12   <2e-16 ***
## data_salud$imc -0.107795   0.007637  -14.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
exp(coefficients(model.logist5))
##    (Intercept) data_salud$imc
##     61.5828259      0.8978114
```

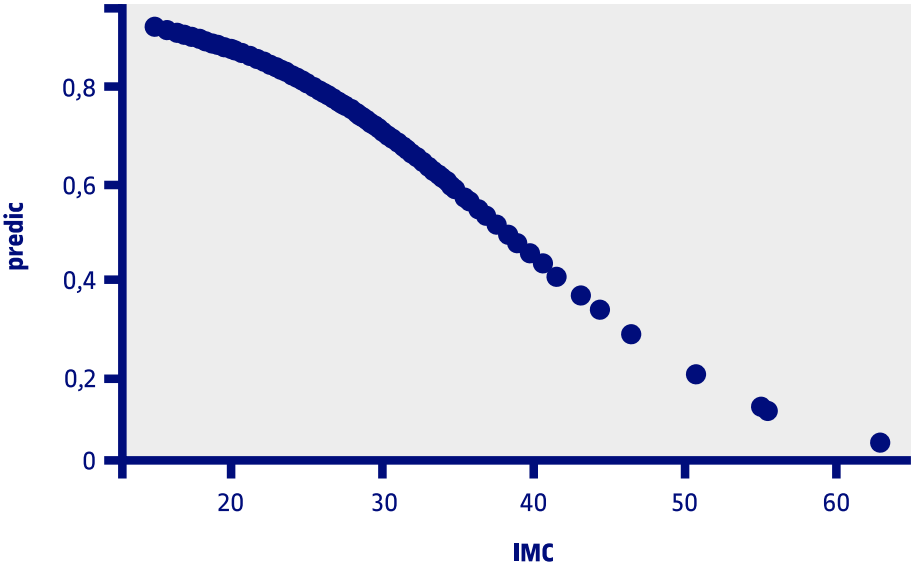
A la vista de los resultados, se puede concluir que, por cada unidad que aumenta el IMC, el *odds* de percibir buena salud es 0,89 veces menor.

Si el IMC aumenta en 10 unidades, el *odds* será $0{,}89^{10} = 0{,}31$ veces menor.

Se pueden ver los resultados en la figura 2.

```
predic=predict(model.logist5,type="response")
plot(imc,predic)
```

Figura 2. Resultados



La curva es decreciente ya que el *odds-ratio* es menor que 1.

Hay que tener cuidado al interpretar los resultados para IMC altos y muy bajos, puesto que están basados en muy pocas observaciones y los resultados obtenidos anteriormente podrían estar sesgados.

4) Caso multivariable: variables independientes continuas y categóricas

Hasta ahora hemos visto cómo ajustar modelos de regresión logística con una sola variable independiente. A estos modelos se les llama univariantes y son válidos en pocas ocasiones.

En la mayoría de los casos, para comprender mejor el fenómeno estudiado, es necesario construir un modelo multivariante. El objetivo es ajustar el efecto de cada variable en el modelo de acuerdo con el efecto del resto de las variables independientes. Por lo tanto, cada coeficiente del modelo proporciona una estimación del logaritmo del *odds* ajustado por las otras variables incluidas en el modelo.

En nuestro ejemplo vamos a tomar las variables independientes: *sexo* y *edad*.

Se tiene la ecuación del modelo:

$$\ln\left(\frac{P(Y=1/X)}{1-P(Y=1/X)}\right)=b_0+b_1x_1+b_2x_2$$

donde: x_1 = sexo y x_2 = edad.

```
model.logist6=glm(formula=salud~sexo+edad,family=binomial(link=logit))
summary(model.logist6)
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.797233   0.120809  31.432  < 2e-16 ***
## sexo2       -0.395887   0.061899  -6.396  1.6e-10 ***
## edad        -0.051446   0.002512 -20.478  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##


exp(coefficients(model.logist6))
## (Intercept)       sexo2        edad
##  44.5776750   0.6730829   0.9498552
```

El parámetro b_1 es muy similar al del modelo univariante. Esto se debe a que la edad media en hombres y mujeres es muy similar (39,15 y 39,32, respectivamente); por lo tanto, al ajustar por edad no cambia el *odds* de la percepción de salud por sexo.

A la hora de interpretar los parámetros del modelo hay que tener en cuenta si se cumple la existencia o no de interacción y confusión.

3.2. Interacción y confusión

La interacción y la confusión son dos conceptos importantes cuando se usan los modelos de regresión multivariantes.

Existe **confusión** cuando la asociación entre la variable respuesta o dependiente y la variable independiente elegida (el factor de interés) difiere significativamente según si se considera, o no, otra variable explicativa. A esta última variable se la denomina *variable de confusión para la asociación*. Es decir, la variable está asociada al mismo tiempo a la variable dependiente o respuesta y al factor de interés. Cuando estas dos asociaciones están presentes, la relación entre la variable respuesta y el factor está confundida.

Existe **interacción** cuando la asociación entre la variable respuesta o dependiente y el factor de interés varía según los diferentes niveles de otra u otras variables explicativas o covariables. Es decir, la covariable modifica el efecto del factor.

Queremos estudiar la existencia de interacción en nuestro ejemplo anterior, teniendo en cuenta las variables independientes *sexo* y *edad* como predictores de la percepción de una buena salud:

```
model.logist7=glm(formula=salud~sexo+edad+sexo:edad,family=binomial(link=logit))
summary(model.logist7)

## glm(formula = salud ~ sexo + edad + sexo:edad, family = binomial(link = logit))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3857   0.3889   0.5161   0.6780   1.1544
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.653489   0.170571  21.419   <2e-16 ***
## sexo2       -0.135450   0.230287  -0.588    0.556
## edad        -0.048200   0.003721 -12.952   <2e-16 ***
## sexo2:edad  -0.005918   0.005046  -1.173    0.241
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Si se observa la tabla de coeficientes por columnas se tiene:

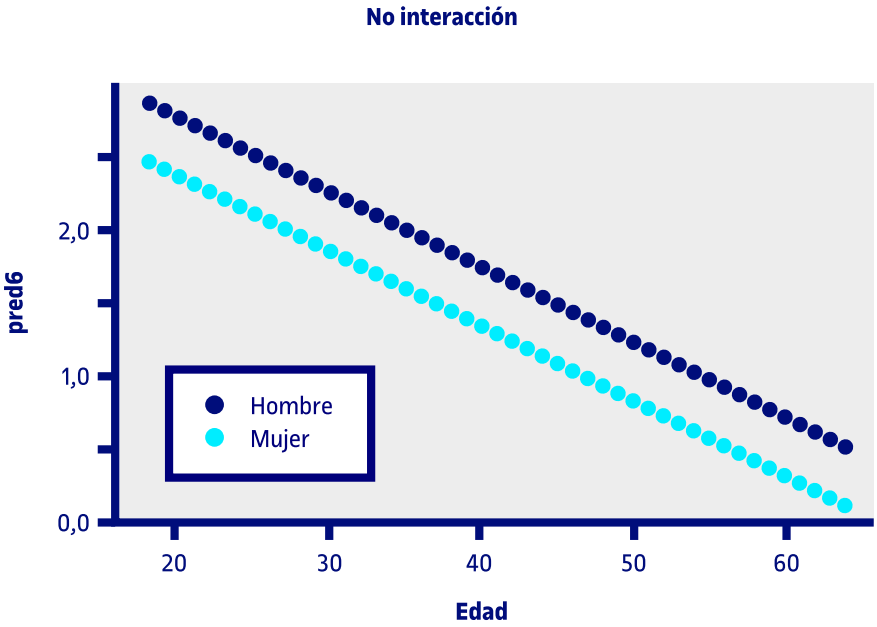
- Estimación de los coeficientes asociados a cada variable (*Estimate*), errores estándar (*Std. error*), valor del estadístico de Wald (*z value*) y el *p-valor* asociado (*Pr(>|z|) = p-value*).
- La variable *sexo* no es significativa. Esta información la proporciona el estadístico de Wald y su *p-valor* asociado (*p-value* = 0,556 > 0,05).
- Además, se concluye que el término de interacción «sexo2:edad» no es estadísticamente significativo (*p-value* = 0,241).

Por otro lado, si representamos el *logit* del modelo, frente a la edad, según la variable *sexo*, se tiene:

```
pred6=predict(model.logist6,type="link")
plot(edad,pred6,type="n",main="NO interaccion")
points(edad[sexo==1],pred6[sexo==1],col=2)
```

```
points(edad[sexo==2],pred6[sexo==2],col=4)
```

Figura 3. Resultados del *logit* del modelo, frente a la edad, según la variable *sexo*



Se tiene que las rectas son paralelas para cada nivel del factor, por lo que la asociación entre la percepción de salud y sexo no varía por la edad, es decir, dicha variable no modifica el efecto del factor sexo. Por lo tanto, no existe interacción.

Un modo de detectar una variable de confusión es ver si los parámetros estimados para el factor de interés cambian sustancialmente al introducir la covariable en el modelo.

Ved también

En el apartado «Selección de variables» se amplía la información del estadístico de Wald.

Este criterio no se puede aplicar al caso de la interacción, ya que la inclusión de un término de interacción, especialmente si una de las variables es continua, normalmente da lugar a cambios en los parámetros estimados. Esto es así aunque la interacción no sea significativa.

En nuestro ejemplo, para saber si la edad es un variable de confusión con respecto al sexo, veremos si el valor del parámetro del sexo cambia mucho al introducir la edad.

Modelo univariante:

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.64320    0.04482  36.661  < 2e-16 ***
## sexo2       -0.37751    0.05990  -6.302 2.93e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.797233   0.120809  31.432  < 2e-16 ***
## sexo2       -0.395887   0.061899  -6.396  1.6e-10 ***
## edad        -0.051446   0.002512 -20.478  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Se puede observar que la diferencia es muy pequeña (–0,37751 y –0,395887), por lo que podemos concluir que no existe confusión.

De acuerdo con los valores obtenidos, la edad no es una variable de confusión ni tampoco es un efecto modificador.

Si en vez de utilizar la covariable *edad* se usa el ***peso***, tenemos:

```
model.logist7b=glm(formula=salud~sexo+peso,family=binomial(link=logit))
summary(model.logist7b)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.712536   0.218700  16.975   <2e-16 ***
## sexo2       -0.825937   0.076422 -10.808   <2e-16 ***
## peso        -0.026188   0.002671  -9.803   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

exp(coefficients(model.logist7b))
## (Intercept)       sexo2        peso
##  40.9575330   0.4378247   0.9741522
```

Al introducir el peso, el parámetro estimado de sexo pasa de –0,377 a –0,825, un descenso de más del 50 %. Por lo tanto, el valor del coeficiente del sexo cambia mucho al introducir el peso.

Además, como se puede comprobar abajo, la variable peso está relacionada con la variable dependiente *percepción de salud* y también con el factor sexo. Todo ello nos indica que el peso es una variable de confusión.

```
logistica1=glm(salud~peso,family=binomial(link=logit))

## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.032881   0.153471  13.246  < 2e-16 ***
## peso        -0.008431   0.002137  -3.946 7.96e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

logistica2=glm(sexo~peso,family=binomial(link=logit))
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 10.437470   0.240520   43.40   <2e-16 ***
## peso        -0.151629   0.003493  -43.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ahora estudiaremos la existencia de interacción. Para ello, se analizará el modelo **con interacción**.

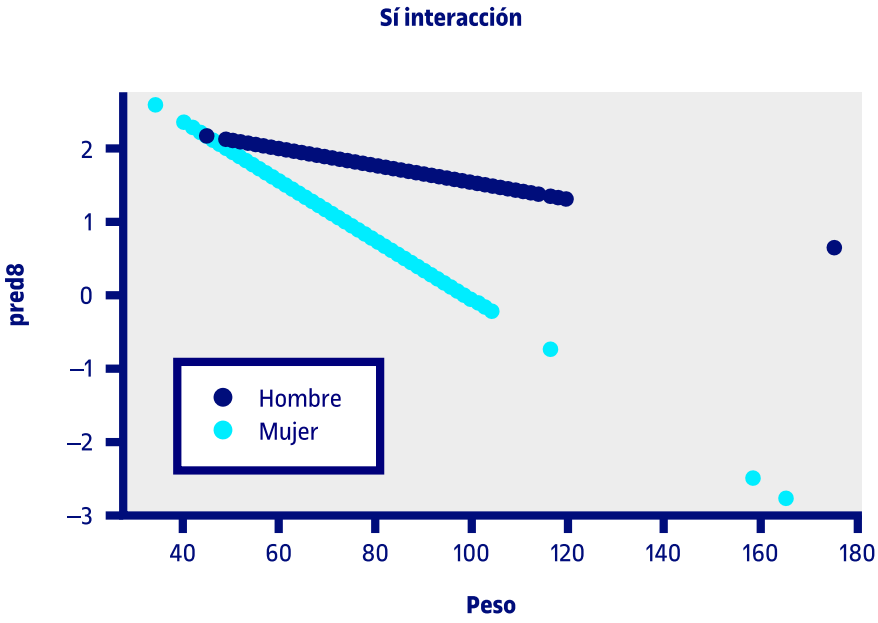
```
model.logist8=glm(formula=salud~sexo+peso+sexo:peso,family=binomial(link=logit))
summary(model.logist8)

## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.561521   0.306033   8.370  < 2e-16 ***
## sexo2        1.236831   0.393019   3.147  0.00165 **
## peso        -0.011713   0.003836  -3.053  0.00226 **
## sexo2:peso  -0.028984   0.005438  -5.330 9.81e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se observa que el estimador de interacción *sexo:peso* es estadísticamente significativo.

De manera análoga al caso anterior, se representará en un gráfico del *logit* del modelo frente al peso, según sea hombre o mujer.

Figura 4. Resultados del *logit* del modelo frente al peso



Se puede observar que las rectas no son paralelas y que la percepción de salud en función del peso es distinta para hombres que para mujeres. Estas se ven más afectadas por el peso.

En resumen, para evaluar si una variable es de confusión, se compara el coeficiente estimado para el factor de riesgo de los modelos que contienen y no contienen la covariable.

Cualquier cambio importante en el coeficiente del factor sugiere que la variable es de confusión. Si esto ocurre, aunque la interacción no sea estadísticamente significativa, la variable ha de salir del modelo. En nuestro caso, se debería eliminar la variable *peso*.

Por otra parte, una variable es un «efecto modificador» solo cuando el término de interacción añadido al modelo es estadísticamente significativo.

Las interacciones no deben eliminarse del modelo, a no ser que alguna de las variables sea de confusión.

3.2.1. Interpretación del OR en presencia de interacción

Antes hemos visto que, cuando existe interacción entre el factor de interés y otra variable, el parámetro estimado para el factor de riesgo depende del valor de la variable que interactúa con este; por lo tanto, no podemos obtener el OR simplemente mediante la exponencial del valor estimado.

Los pasos que hay que seguir se detallan a continuación:

1) Escribir la ecuación del *logit* para los dos niveles del factor de riesgo.
2) Calcular la diferencia entre los *logit*.
3) Calcular la exponencial del valor obtenido.

Por ejemplo, consideremos un modelo que contiene solo dos variables y su interacción. Llamamos *f* al factor, *x* a la covariable y $f * x$ a la interacción.

$$\ln\left(\frac{P(Y=1/X)}{1-P(Y=1/X)}\right) = b_0 + b_1 f + b_2 x + b_3 f * x$$

Se calcula el OR para los niveles de *f*, que llamaremos f_1 y f_0.

$$\ln 1\left(\frac{P(Y=1/X)}{1-P(Y=1/X)}\right) = b_0 + b_1 f_1 + b_2 x + b_3 f_1 * x$$

$$\ln 0\left(\frac{P(Y=1/X)}{1-P(Y=1/X)}\right) = b_0 + b_1 f_0 + b_2 x + b_3 f_0 * x$$

Se resta:

$$\ln 1\left(\frac{P(Y=1/X)}{1-P(Y=1/X)}\right) - \ln 0\left(\frac{P(Y=1/X)}{1-P(Y=1/X)}\right) = b_1(f_1 - f_0) + b_3 x(f_1 - f_0)$$

Por lo tanto:

$$\mathrm{OR} = \exp\left((b_1(f_1 - f_0) + b_3 x(f_1 - f_0)\right)$$

Si lo aplicamos a nuestro ejemplo anterior, tomando la interacción *sexo*peso*, se tiene:

$$\mathrm{OR} = \exp(1{,}23 - 0{,}029x).$$

3.3. Selección de variables

Del conjunto de las variables que tenemos en nuestro estudio, ¿cuáles deben introducirse en el modelo?

Se busca el modelo más sencillo que ajuste bien los datos. Hay que tener en cuenta que un mayor número de variables en el modelo implicará mayores errores estándar.

En el modelo final deben considerarse también los términos de interacción que se van a introducir.

Las **estrategias** de selección de variables más usadas son secuenciales:

- **Selección hacia adelante**, en la cual en cada etapa se añade la mejor variable predictora aún no seleccionada.

- **Eliminación hacia atrás**, en la cual partiendo del conjunto completo de variables predictoras se va eliminando en cada etapa la peor, hasta que las variables que quedan en el modelo son todas ellas pertinentes.

- **Modelización paso a paso**, en la cual se combinan las dos estrategias anteriores.

Una vez elegida la estrategia, tenemos un modelo preliminar y se debe determinar si las variables incluidas son significativas o no. Esto supone la formulación de test de hipótesis, o bien cuando proceda la realización del contraste de independencia chi-cuadrado, para determinar si las variables explicativas o independientes del modelo están relacionadas significativamente con la variable dependiente o respuesta.

Se pueden hacer contrastes no solo sobre cada coeficiente estimado, sino también para comparar modelos.

1) Test de Wald: contraste de hipótesis de nulidad de los parámetros

Se aplica a cada una de las variables del modelo y tan solo puede ser usado para testar un único parámetro.

Se basa en contrastar si es cero o no un determinado coeficiente estimado b_i, para la variable x_i.

$$\mathrm{H}_0 = b_i = 0$$
$$\mathrm{H}_1 = b_i \neq 0$$

Donde el estadístico del contraste sigue una distribución $N(0, 1)$.

$$z = \frac{b_i}{\hat{S}_{b_i}}$$

Aplicado a nuestro ejemplo y tomando el `model.logist1`, se tiene el resumen siguiente:

```
model.logist1=glm(formula=salud~sexo+factor(bebedor_Rel),family=binomial(link=logit))
summary(model.logist1)
##
## Call:
## glm(formula = salud ~ sexo + factor(bebedor_Rel), family = binomial(link = logit))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9851   0.5480   0.5629   0.7040   0.7814
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            1.26858    0.06027  21.049  < 2e-16 ***
## sexo2                 -0.23868    0.06230  -3.831 0.000128 ***
## factor(bebedor_Rel)1  0.55151    0.06288   8.771  < 2e-16 ***
## factor(bebedor_Rel)2  0.49377    0.15984   3.089 0.002007 **
```

donde se presentan la tabla de coeficientes (*Estimate*), errores estándar (*Std. error*), valor del estadístico de Wald (*z value*) y el *p-valor* asociado (*Pr(>|z|)*). Se puede observar que el *p-valor* es menor de 0,05, por lo que las variables elegidas son significativas.

2) AIC: criterio de información de Akaike

Un criterio comúnmente utilizado para determinar qué variables deben introducirse en el modelo es el llamado *AIC*.[2]

[2]Acrónimo del inglés, *Akaike information criterion*.

En algunas ocasiones no sabemos qué variables son más relevantes en el modelo. En el caso de regresión logística se considera la variable más importante aquella que da lugar al cambio más grande en el índice AIC, el cual está basado en el *deviance* y el número de parámetros del modelo. Primero se ajusta el modelo con solo la constante y se compara con el modelo que resulta de introducir cada una de las variables. Aquella variable que da lugar a un valor del AIC más pequeño es considerada la más importante y se introduce en el modelo.

A continuación, se compara el modelo inicial con una variable con el modelo posterior que resulta de introducir otra variable más y así sucesivamente. Además, en cada nuevo paso se comprueba que las variables que se han introducido en el paso anterior siguen siendo significativas al introducir una nueva variable. Una vez que se han elegido todas las variables importantes, hay que comprobar las interacciones entre estas. Recordemos que la interacción entre dos variables significa que el efecto de una de ellas no es constante para todos los niveles de la otra. La decisión final sobre si un término de interacción debe ser incluido en el modelo debe basarse en consideraciones tanto estadísticas como científicas. Seleccionadas las variables y sus interacciones, habrá que comprobar los modelos con los test de bondad del ajuste, que se verán posteriormente.

Volviendo al ejemplo visto en el apartado «Modelos de regresión logística binaria», se tenía el modelo siguiente:

```
model.logist1=glm(formula=salud~sexo+factor(bebedor_Rel),family=binomial(link=logit))
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.26858    0.06027  21.049  < 2e-16 ***
## sexo2                -0.23868    0.06230  -3.831 0.000128 ***
## factor(bebedor_Rel)1  0.55151    0.06288   8.771  < 2e-16 ***
## factor(bebedor_Rel)2  0.49377    0.15984   3.089 0.002007 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7177.9  on 7356  degrees of freedom
## Residual deviance: 7059.2  on 7353  degrees of freedom
## AIC: 7067.2
##
## Number of Fisher Scoring iterations: 4
```

En la salida de R aparece el AIC bajo la *residual deviance*: **AIC = 7.067,2**.

Si lo comparemos con el obtenido al tomar el modelo solo con la variable independiente sexo:

```
model.logist2=glm(formula=salud~sexo,family=binomial(link=logit))
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.64320    0.04482  36.661  < 2e-16 ***
## sexo2       -0.37751    0.05990  -6.302 2.93e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7177.9  on 7356  degrees of freedom
## Residual deviance: 7137.8  on 7355  degrees of freedom
## AIC: 7141.8
```

Se obtiene un **AIC de 7.141,8**, mayor al calculado anteriormente, por lo que el primer modelo ajusta mejor los datos.

3.4. Bondad del ajuste

Se dice que un modelo presenta un buen ajuste a los datos si los valores que predice reflejan de manera adecuada los valores observados. Si el modelo presenta un mal ajuste, este no puede ser utilizado para extraer conclusiones ni efectuar predicciones.

Un modo de medir la adecuación de un modelo es proporcionando medidas globales de **bondad de ajuste** mediante test estadísticos.

Existen varias medidas de ajuste global para comparar la diferencia entre valores predichos y valores observados. Tres de las más utilizadas son el test basado en la *devianza* D, el estadístico χ^2 de Pearson y el test de Hosmer-Lemeshow. Los dos primeros se basan en los patrones de las covariables y pueden ser usados en los modelos lineales generalizados (MLG) en general. El tercero se basa en probabilidades estimadas y se aplica en el caso de un MLG con distribución binomial, es decir, un modelo de regresión logística.

3.4.1. Test basado en la *devianza* D. Test de razón de verosimilitudes

Este test tiene como objetivo comparar dos modelos de regresión logística, el denominado modelo completo (*full model*) y el que se conoce como modelo reducido (*reduced model*).

La hipótesis nula testada en el test establece que los parámetros correspondientes a las variables que forman parte del modelo completo, pero no del modelo reducido, valen cero.

Para comprender lo dicho anteriormente, consideramos tres modelos expresados en su formulación *logit*:

1) Modelo 1: *logit* $1 = b_0 + b_1x_1 + b_2x_2$
2) Modelo 2: *logit* $2 = b_0 + b_1x_1 + b_2x_2 + b_3x_3$
3) Modelo 3: *logit* $3 = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_1x_3 + b_5x_2x_3$

Tal y como puede verse, el modelo 2 es una extensión del modelo 1, de igual modo que el modelo 3 lo es del 2.

En caso de querer comparar el modelo 2 con el modelo 1, este último desempeñaría el papel de «modelo reducido», mientras que el modelo 2 sería el «modelo completo». De manera análoga, si quisiésemos comparar el modelo 3 con el modelo 2, dicho modelo 2 sería el «modelo reducido», mientras que el modelo 3 sería el «modelo completo».

Supongamos que queremos comparar el modelo 2 con el 1.

El **test de la razón de verosimilitud** plantea como hipótesis nula que $b_3 = 0$, es decir, que el parámetro del componente que forma parte del modelo 2, pero no del modelo 1, es cero.

Por lo tanto, se tiene:

$$H_0: b_3 = 0$$
$$H_1: b_3 \neq 0$$

El estadístico del contraste se construye a partir de la diferencia de *devianza* entre los modelos que queremos comparar.

Se define ***devianza*** del modelo (estadístico D) una medida del grado de diferencia entre las frecuencias observadas y las predichas por el modelo de la variable dependiente. Dicho estadístico mide en qué grado el modelo se ajusta a los datos. Cuanto menor sea su valor, mejor es el ajuste.

Se calcula a partir de la transformación logarítmica de la función de verosimilitud del modelo (***log likelihood function***).

Se define la función de verosimilitud de la siguiente manera:

$$L(\beta) = p_i^{\sum y_i} * (1 - p_i)^{n - \sum y_i}$$

donde:

- $P(Y_i = 1) = p_i$
- $P(Y_i = 0) = (1 - p_i)$
- n: tamaño de la muestra

Dada la transformación logarítmica:

$$\ln L(\beta) = \sum y_i * \ln (p_i) + \left(n - \sum y_i \right) * \ln (1 - p_i)$$

Se tiene: ***devianza* = D = –2 ln $L(\beta)$**.

Para comprobar si una variable mejora el modelo, se compara el valor de D con y sin la variable independiente incluida en el modelo, usando el **test de la razón de verosimilitud**. Este test utiliza el estadístico G, que se construye a partir de la diferencia de *devianza* entre los modelos que queremos comparar.

G = D (modelo sin la variable) – D (modelo con la variable),

o lo que es lo mismo,

D (modelo reducido) – D (modelo completo)

El estadístico G así construido tiene una distribución asintótica chi-cuadrado, con r grados de libertad igual al número de parámetros que, en el modelo completo, deben igualarse a cero para que dicho modelo completo coincida con el modelo reducido.

- Si se compara el modelo 2 con el 1, los grados de libertad serían $r = 1$.
- Si se compara el modelo 3 con el 2, los grados de libertad serían $r = 2$.

Por lo tanto, la hipótesis nula será rechazada para el nivel de significación α cuando $D \geq \chi^2_r$, que es equivalente a que el *p-valor* del contraste sea menor que el nivel α fijado.

Retomando el ejemplo sobre la percepción de la salud, tomando como variable independiente sexo, se tenía el resumen del modelo (*summary*) siguiente:

```
model.logist2=glm(formula=salud~sexo,family=binomial(link=logit))
summary(model.logist2)
##
## Call:
## glm(formula = salud ~ sexo, family = binomial(link = logit))
##
## Null deviance: 7177.9  on 7356  degrees of freedom
## Residual deviance: 7137.8  on 7355  degrees of freedom
```

El valor de la *devianza* para el modelo solo incluye la constante (*null deviance*) seguida de sus grados de libertad, y el valor de la *devianza* para el modelo ajustado (*residual deviance*) y sus grados de libertad. Para que el modelo sea bueno, la *devianza* residual debe ser menor que la *devianza* nula. En este caso la *null deviance* es 7.177,9, pero cuando se incluye la variable *sexo*, la *residual deviance* es 7.137,8, lo que nos dice que con esta variable el modelo mejora.

Para poder valorar mejor la eficacia del modelo, se analizará el contraste de la razón de verosimilitud. El estadístico G que mide la diferencia entre las dos *devianza* (del modelo con y sin incluir la variable), será:

$$G = 7.177{,}9 - 7.137{,}8 = \mathbf{40{,}113}$$

El programa R lo calcula de la manera siguiente:

```
anova(model.logist2,test="Chisq")
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: salud
##
## Terms added sequentially (first to last)
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                  7356     7177.9
## sexo  1   40.113      7355     7137.8 2.397e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

3.4.2. Estadístico chi-cuadrado de Pearson

Este estadístico χ^2 puede calcularse como la suma de los cuadrados:

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

donde r_j son los *Pearson residuals* (residuos de Pearson).

$$r_j = \frac{y_j - m_j * p_j}{\sqrt{m_j * p_j * (1 - p_j)}}$$

donde:

- m_j: número de individuos que comparten el mismo patrón de covariables, es decir, la misma combinación de variables explicativas.
- J: número de patrones diferentes.

El estadístico χ^2 tiene la misma distribución asintótica que la de la diferencia de *devianza* entre los modelos que queremos comparar (estadístico G), es decir, una chi-cuadrado con los mismos grados de libertad.

Por lo tanto, la hipótesis nula será rechazada para el nivel de significación α cuando $D \geq \chi^2_r$.

Para efectuar este análisis con el programa R, primero se calcula el estadístico como la suma de cuadrados de los residuos de Pearson y después la significación del test.

Como se aprecia, el test sigue siendo significativo para el modelo anterior.

```
sum(residuals(model.logist2,type="pearson")^2)
## [1] 7357
1-pchisq(sum(residuals(model.logist2,type="pearson")^2),1)
## [1] 0
```

3.4.3. Test de Hosmer-Lemeshow

Si una de las variables explicativas es continua, no deben usarse los test descritos anteriormente, sino el test de Hosmer-Lemeshow.

Este test consiste en comparar los valores previstos (esperados) por el modelo con los valores observados.

Ambas distribuciones, esperada y observada, también se contrastan mediante una prueba χ^2.

La **hipótesis nula (H_0)** del test indica que no hay diferencias entre los valores observados y los valores pronosticados, por lo tanto, el rechazo de H_0 indicaría que el modelo no está bien ajustado.

El valor observado es el dato que tenemos, mientras que el esperado es el valor esperado teórico calculado mediante el modelo construido.

Aunque no vamos a explicar aquí los fundamentos matemáticos, conceptualmente este test consiste en dividir el recorrido de valores de la variable dependiente dicotómica *Y* en una serie de intervalos. Estos intervalos deben contener un número de observaciones suficientemente grande (cinco o más). Se trata de contar intervalo por intervalo el valor esperado y el observado para cada uno de los dos resultados posibles de la variable dependiente dicotómica.

Función para el test de Hosmer-Lemeshow

En la librería Resource Selection hay una función que ajusta este test.

Se tomará como ejemplo el modelo con la variable continua *IMC* y la variable categórica *sexo*.

```
model=glm(formula=salud2~sexo+imc,family=binomial(link=logit))
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.968533   0.214406   23.17   <2e-16 ***
## sexo2       -0.649257   0.063516  -10.22   <2e-16 ***
```

```
## imc         -0.128286   0.007951  -16.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
library(ResourceSelection)
hoslem.test(salud2,fitted(model))

##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  salud2, fitted(model)
## X-squared = 12.129, df = 8, p-value = 0.1455
```

El *p-valor* es de 0,1455, lo que indicaría que el modelo ajusta bien los datos.

Cada vez que se incluye o elimina una variable hemos de comparar el nuevo modelo con el anterior hasta que todas las variables importantes estén en el modelo final.

Un modelo de regresión logística puede resultar inadecuado por diferentes razones.

Una de ellas es la invalidez del componente lineal del modelo. Esta situación es frecuente cuando ciertas variables explicativas o términos de interacción no son incluidos en el modelo cuando deberían serlo, o también cuando no se realizan transformaciones en ellas que permitan mejorar el ajuste a los datos. La existencia de *outliers* puede ser también determinante en un mal ajuste.

Recordad que el modelo logístico no asume **linealidad** entre la variable dependiente y una covariable, pero sí respecto al *logit* de la *odss* de la variable dependiente.

Si no se cumpliera esta asunción, la relación entre dicha covariable y la variable dependiente estaría infraestimada.

Una manera de comprobar si alguna de las covariables no cumple esta asunción de linealidad respecto al *logit* es dibujar los gráficos de dispersión.

3.5. Predicciones del modelo

Una vez estimados los coeficientes del modelo, es posible conocer la probabilidad de que la variable dependiente elegida tome el valor 1, en presencia de las variables independientes X.

$$P(Y=1/X)=\frac{e^{b_0+\sum_{i=1}^{n}(b_i x_i)}}{1+e^{b_0+\sum_{i=1}^{n}(b_i x_i)}}$$

Para calcular dichas probabilidades en R, se usa el comando `predict( )`.

Por ejemplo, se quiere predecir la probabilidad de percepción de buena salud en el caso de ser mujer y tener un IMC de 28.

```
model=glm(formula=salud2~sexo+imc,family=binomial(link=logit))
pred<-predict(model, data.frame(sexo="1",imc=28),type ="response")
pred
##         1
## 0.7984337
```

De acuerdo con el modelo, se obtiene una probabilidad de 0,8.

Además, cada vez que se ejecuta el comando `glm( )`, para cada individuo de la muestra se obtiene un valor de la probabilidad predicha por el modelo. Estas probabilidades quedan almacenadas en un vector de R que es el componente `fitted.values` del modelo.

```
head(model$fitted.values)
##         1         2         3         4         5         6
## 0.8595336 0.8069657 0.8642668 0.8613820 0.8648676 0.8037482
```

Una de las principales aplicaciones de un modelo logístico es clasificar las observaciones en función del valor que tome el predictor, es decir, la ecuación del modelo nos proporciona una probabilidad que nos permite predecir a partir de ella para cada sujeto un valor de la variable dependiente Y, que llamaremos Y_{pred} (Y predicho).

Para conseguir esta clasificación, es necesario establecer un umbral (*threshold*) de probabilidad a partir del cual se considerará que Y_{pred} tomará el valor 0 o 1.

Por ejemplo: se considera como umbral una probabilidad de 0,05 y se crean dos grupos según si el valor predicho está por encima o por debajo de aquel. Entonces:

- Sujetos cuya $P(Y = 1/X) > 0{,}05$, se clasificarían en grupo 1, donde $Y_{\text{pred}} = 1$.
- Sujetos cuya $P(Y = 1/X) \leq 0{,}05$, se clasificarían en grupo 2, donde $Y_{\text{pred}} = 0$.

3.5.1. Matriz de confusión

Como hemos visto anteriormente, la ecuación del modelo nos proporciona una probabilidad ajustada por el modelo que nos permite predecir a partir de ella para cada sujeto un valor de la variable dependiente Y, que llamaremos Y_{pred} (Y predicho).

Para construir la **matriz de confusión** utilizamos los casos del conjunto de datos original (que puede denominarse *conjunto de aprendizaje* o *training set*). Para cada caso se tiene una respuesta observada y se puede predecir la probabilidad ajustada por el modelo.

La matriz de confusión permite comparar estos dos valores. Por ejemplo, un individuo en el conjunto de datos puede tener una respuesta igual a 1, y la probabilidad estimada por el modelo puede ser 0,9, o lo que es lo mismo, un 90 %. Ese caso es bastante plausible. Pero puede ocurrir que la respuesta sea 1, mientras que la predicción del modelo sea muy baja, digamos del 20 %. En este caso, la predicción del modelo sería sorprendente.

En el ejemplo del fabricante de perfume, visto en el apartado «Modelos lineales generalizados (MLG)», consideremos a una persona que elige comprar la fragancia pero que tiene unas características totalmente alejadas de los consumidores más habituales. Eso le daría

una probabilidad estimada de compra de solo el 20 %. Podemos decir que la probabilidad de que compre el producto es de 1 contra 4.

En la mayoría de los conjuntos de datos cuando se comparan predicciones y observaciones, podemos encontrar casos concretos en los que la observación y la probabilidad previstas no son concordantes. Esos son errores del modelo. Puede suceder que se dé el caso contrario al que se acaba de comentar, una muy alta probabilidad de elegir la respuesta afirmativa; sin embargo, se observa lo contrario.

En modelos de regresión lineal, era razonable discutir la correlación entre observaciones y predicciones en términos de estadístico R^2, el coeficiente de determinación. Sin embargo, en modelos generalizados las correlaciones carecen del mismo sentido.

En un modelo de regresión logística, cuando se observa la probabilidad predicha por el modelo no se sabe si el valor es demasiado alto o demasiado bajo. Normalmente un buen umbral de discriminación es el del 50 %. Eso corresponde a un *odds* igual a 1, lo que significa que las dos opciones de respuesta son equiprobables. Fijado dicho umbral se considera que si la respuesta es positiva y la probabilidad estimada es superior al 50 %, entonces el modelo ha acertado.

De igual modo, si la respuesta es negativa y la probabilidad es inferior al 50 %, el modelo también ha logrado acertar. Sin embargo, cuando la probabilidad no se corresponde con lo observado, se trata de casos en los que el modelo no acierta.

Veamos un ejemplo de tabla de confusión:

Tabla 3. Ejemplo de matriz de confusión. Número de casos

	Probabilidad predicha inferior al 50 %	**Probabilidad predicha superior o igual al 50 %**	**Total**
Observado: Sí	50	10	60
Observado: No	30	60	90
Total	80	70	150

Del total de 150 casos, vemos que 110 están correctamente clasificados, ya que el modelo predice la respuesta correcta para los 50 participantes que sí comprarían el producto y tienen una probabilidad predicha elevada, y para los 60 que no lo comprarían y tienen una probabilidad baja. Esto significa que el porcentaje general de clasificación correcta es igual a 110/150 = 73,33 %, lo cual es excelente. Sin embargo, hay un número de casos que no tienen la respuesta esperada. Por ejemplo, hay 30 casos que no eligen «comprar el producto», pero la probabilidad es alta y el modelo predice que es probable que hayan hecho esta elección. Por otro lado, hay 10 casos que responden que comprarían el producto pero el modelo predice que su probabilidad de compra es inferior al 50 %.

Para que un modelo de regresión logística con fines predictivos tenga éxito, el número de casos que se clasifican correctamente tiene que ser alto, mientras que el número de casos que se clasifican incorrectamente debe ser bajo.

El número de resultados denominados *falsos positivos* corresponde a casos en los que la predicción de la probabilidad de la respuesta afirmativa es elevada, pero la respuesta observada es negativa. En este ejemplo existen 30 falsos positivos. Los falsos positivos no son buenos en la práctica.

En nuestro ejemplo, es posible que se haya hecho una campaña publicitaria para llegar a las personas que tienen intención de comprar; sin embargo, no se produce la respuesta deseada. Los falsos positivos pueden conducir, en este ejemplo, al fracaso de los esfuerzos comerciales. Si usamos el modelo para predecir el comportamiento de estos clientes y tratamos de venderles la «nueva fragancia» sin éxito, incurriremos en unos costes publicitarios que podrían ser ahorrados.

El número de respuestas correspondientes a los falsos negativos corresponde al número de casos donde el modelo predice que el cliente tiene una probabilidad de compra baja; sin embargo, los participantes sí han elegido comprar. En nuestro ejemplo, los clientes que han elegido comprar la fragancia ante una predicción de hacerlo baja, inferior al 50 %, corresponden a un total de 10 casos. Este modelo tiene solo unos pocos falsos negativos. Y si se utilizara con fines predictivos, significa que solo habría unos cuantos clientes que, aun no respondiendo al perfil del comprador, comprarían el producto.

Idealmente, un modelo perfecto en términos de capacidad predictiva no tendría ni falsos positivos ni falsos negativos en la tabla de confusión. Pero existen otras medidas que también suelen tenerse en cuenta. La **sensibilidad** es la proporción de los clasificados correctamente entre los verdaderos participantes que han dado respuesta afirmativa. La **especificidad** es la proporción de casos correctamente clasificados entre las respuestas negativas. En la tabla 3, la sensibilidad es 50/(50 + 10) = 83,33 % y la especificidad es 60/(60 + 30) = 66,67 %.

3.5.2. Curva ROC (*receiver operating characteristic*)

La curva ROC es un gráfico de la sensibilidad frente a 1 menos la especificidad. Cada punto en la curva corresponde a un nivel umbral de discriminación en la matriz de confusión. Es decir, así como en el ejemplo anterior se había considerado un umbral del 50 %, se construyen todas las matrices cambiando dicho umbral desde el 1 hasta el 99 %, y se va calculando la sensibilidad y 1 menos la especificidad. El mejor modelo en términos de ajuste sería aquel

modelo que tuviera una curva ROC lo más cerca posible de la esquina superior izquierda de la gráfica. Un modelo no discriminante tendría una curva ROC plana, cerca de la diagonal.

El **análisis ROC** proporciona un modo de seleccionar modelos posiblemente óptimos y subóptimos basado en la calidad de la clasificación a diferentes niveles o umbrales. Para tener una regla objetiva de comparación de las curvas ROC, se calcula el área bajo la curva, simplemente llamada AUROC (*area under the ROC*). El modelo cuya área sea superior es el preferido.

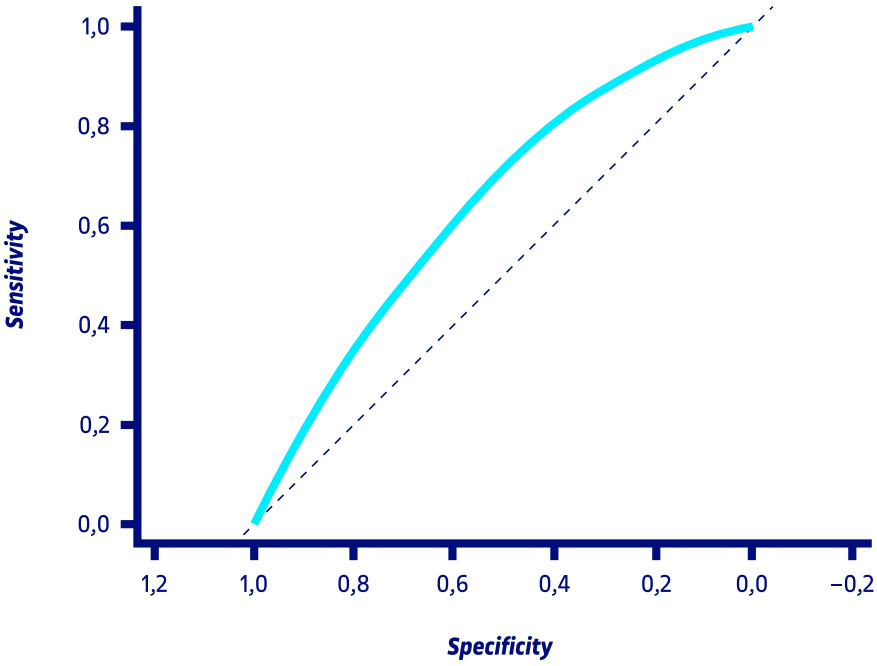
En general:

- Si AUROC ≤ 0,5, el modelo no ayuda a discriminar.
- Si 0,6 ≤ AUROC < 0,8, el modelo discrimina de manera adecuada.
- Si 0,8 ≤ AUROC < 0,9, el modelo discrimina de forma excelente.
- Si AUROC ≥ 0,9, el modelo discrimina de modo excepcional.

En el ejemplo sobre la percepción de la salud podemos visualizar la curva ROC de la siguiente manera:

```
model=glm(formula=salud2~sexo+imc,family=binomial(link=logit))
library(pROC)
prob=predict(model, data_salud, type="response")
r=roc(salud,prob, data=data_salud)
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
plot (r)
```

Figura 5. Curva ROC




```
auc(r)
## Area under the curve: 0.6412
```

A la vista de los resultados obtenidos, el modelo no discrimina muy bien los datos.

Uno de los elementos prácticos más importantes es saber seleccionar el **umbral de clasificación** (*threshold*) más adecuado para que el modelo sea útil desde un punto de vista predictivo. Eso quiere decir que, a lo mejor, no es lo más adecuado fijar un punto de corte en el 50 %. Para ello, conviene fijarse en aquel punto que proporciona una mayor distancia entre la curva ROC y la diagonal. Ese nivel de umbral de clasificación es el que da una mayor sensibilidad y especificidad en el modelo, es decir, una mayor capacidad de clasificar correctamente a los participantes en el estudio.

4. Otros modelos lineales generalizados

Los modelos lineales generalizados (MLG) cubren los modelos estadísticos más utilizados, como la regresión lineal para las respuestas distribuidas normalmente, el modelo logístico para datos binarios y los modelos *probit* y de Poisson.

A continuación se explican los dos modelos, modelo *probit* y modelo de Poisson.

1) Modelo *probit*

El modelo *probit* se utiliza para casos en los que la respuesta es dicotómica y se quiere una alternativa al modelo de regresión logística. Aunque existe una equivalencia entre los parámetros de ambos modelos demostrada por Amemiya (1981) –si se dividen por 1,6 los coeficientes del modelo *logit*, se obtienen los del modelo *probit*–, todavía hay veces en las que se prefiere el modelo *logit* porque es mejor para casos en los que existen más observaciones extremas y se quiere insistir en la interpretación de los *odds*.

En el modelo *probit*, la especificación es la siguiente:

$$\text{Prob}(Y_i = 1) = \int_{-\infty} \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \ldots + \beta_k Xk_i \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{u2}{2}\right)} du$$

2) Modelo de Poisson

El modelo de Poisson se utiliza para casos en los que la variable dependiente es el número de veces en las que ocurre algún suceso. Se especifica como sigue:

$$E(Y_i) = \exp(\beta_0 + \beta_1 X1_i + \beta_2 X2_i + \ldots + \beta_k Xk_i)$$

En este caso se supone que la distribución de probabilidad de Y_i es una distribución de Poisson. La interpretación de los parámetros se realiza en términos de su signo. Cuando un parámetro es positivo, eso significa que si su característica asociada aumenta, entonces también lo hace el número esperado de veces que tiene lugar el fenómeno analizado. En caso contrario, si el parámetro es negativo, entonces al aumentar dicha variable disminuye el número esperado que se está modelizando.

Un ejemplo concreto de modelo de Poisson es el de ver cuántas noches pernoctan los turistas que acuden a una ciudad. Si se obtiene que la edad tiene un coeficiente positivo, entonces se espera que a mayor edad más elevado sea el número de noches que se espera que pasen los turistas en la ciudad. Si, por ejemplo, el coeficiente asociado a una nacionalidad fuera negativo, se interpretaría que los turistas de dicha nacionalidad se espera

que pasen menos noches en la ciudad que los que provienen de otros países, suponiendo que tuvieran el resto de las características iguales.

Resumen

En este módulo se ha presentado la regresión logística (RL) como parte del conjunto de métodos estadísticos denominados *modelos lineales generalizados* (MLG).

La RL es adecuada cuando la variable de respuesta *Y* es politómica (admite varias categorías de respuesta, tales como «empeora mucho», «empeora», «se mantiene», «mejora», «mejora mucho»), pero es especialmente útil cuando la variable de respuesta es dicotómica, es decir cuando existen dos posibles respuestas.

La RL es una técnica de regresión muy empleada en la producción científica contemporánea. Contesta a preguntas tales como: ¿se puede predecir con antelación si un cliente que solicita un préstamo a un banco va a ser un cliente moroso?, ¿una empresa va a entrar en bancarrota?, ¿se puede predecir de antemano que un paciente corra riesgo de padecer una enfermedad?, ¿es útil una campaña publicitaria?, etc.

A lo largo de este material se ha presentado la RL como método de análisis aplicado a la investigación basada en un conjunto de datos. Para ello, se ha resuelto un caso práctico con datos reales sobre la percepción del estado de salud, como hilo conductor. En primer lugar se presentan los conceptos teóricos fundamentales, como la ecuación del modelo, la función logística, la noción de *odds* y *odds-ratio*, el *logit*. Posteriormente, se explica cómo estimar los coeficientes del modelo y se va resolviendo el caso práctico, tomando primero un modelo con variables sencillas, para luego aplicarlo a tipo y número de variables más complejas. Por último, se explican los métodos de bondad y ajuste del modelo, así como los métodos de predicción.

La identificación del mejor modelo de regresión logística se realiza mediante la comparación de modelos utilizando el test de razón de verosimilitud, que indica, a partir de los datos de la muestra, cuánto más probable es un modelo frente al otro. Un criterio comúnmente utilizado para determinar qué variables deben introducirse en el modelo es el llamado *AIC* (*akaike information criterion*).

En cada caso, se muestran los principales estadísticos que ofrece la salida del programa R, así como las conclusiones obtenidas de los mismos.

Ejercicios de autoevaluación

1. Se realiza un modelo de regresión logística donde la variable que se codifica como 1 es padecer una enfermedad y como 0 el no padecerla. ¿Qué nos indica que un determinado factor tenga OR de 1,3?

a) Que estamos ante un factor de protección.
b) Que estamos ante un factor de riesgo.
c) No influye.
d) Estamos ante un factor de riesgo pero siempre y cuando la variable sea significativa.
e) Ninguna de las respuestas anteriores es correcta.

2. Se realiza un modelo de regresión logística donde la variable que se codifica como 1 es padecer una enfermedad y como 0 el no padecerla. ¿Qué nos indica que un determinado factor significativo tenga un OR de 0,5?

a) Que estamos ante un factor de protección.
b) Que estamos ante un factor de riesgo.
c) No influye.
d) Las OR deben ser mayores que 1.
e) Ninguna de las respuestas anteriores es correcta.

3. El contraste que emplea la regresión logística para ver si las variables son significativas es...

a) el test T.
b) el test de Wald.
c) el coeficiente de Spearman.
d) el coeficiente de correlación lineal.
e) Ninguna de las respuestas anteriores es correcta.

4. ¿Por qué es útil el modelo de regresión logística?

a) Porque permite estimar los factores de riesgo del fenómeno estudiado.
b) Porque permite comprobar si existe relación entre variables.
c) Porque permite usar tanto variables cualitativas como cuantitativas.
d) Todo lo anterior.
e) Ninguna de las respuestas anteriores es correcta.

5. Se quiere valorar si un modelo ajusta o no bien los datos. Dicho modelo contiene factores continuos y cualitativos. ¿Qué test elegiríais?

a) El gráfico Classplot.
b) El test de Hosmer-Lemeshow.
c) El test de Wald.
d) El test de razón de verosimilitud.
e) El test de Kruskall-Wallis.

6. Si al determinar el modelo obtenemos un 60 % de *outliers*, nuestra impresión será la siguiente:

a) El modelo es bueno.
b) El modelo no es bueno.
c) No tiene nada que ver con la calidad del modelo.
d) Nada de lo anterior.
e) Todas las respuestas anteriores son correctas.

7. Si tuvierais que decidir qué variables influyen más en el éxito de una campaña publicitaria, ¿qué técnica elegiríais?

a) El análisis de anova.
b) El contraste de hipótesis.
c) El modelo de regresión logística, ya que este incluye contrastes y nos permite además modelizar y predecir.

d) El análisis clúster.
e) Ninguna de las respuestas anteriores es correcta.

8. La estimación de los parámetros del modelo se realiza mediante el método de...

a) máxima verosimilitud.
b) mínimos cuadrados.
c) máximos cuadros.
d) mínimos cuadrados ponderados.
e) Ninguna de las respuestas anteriores es correcta.

9. Las variables independientes de un modelo de regresión logística pueden ser...

a) continuas.
b) dicotómicas.
c) categóricas.
d) Todas las anteriores.
e) Ninguna de las respuestas anteriores es correcta.

10. El análisis *probit* es un modelo conceptualmente similar a...

a) el análisis discriminante.
b) el análisis factorial.
c) la regresión logística.
d) Todas las anteriores.
e) Ninguna de las respuestas anteriores es correcta.

Solucionario

Ejercicios de autoevaluación

1. d

2. a

3. b

4. d

5. b

6. b

7. c

8. a

9. d

10. c

Bibliografía

Amemiya, T. (1981). «Qualitative response models: A survey». *Journal of Economic Literature* (vol. 19, n.º 4, págs. 1483-1536).

Artís, M.; Clar, M.; Barrio, T.; Guillén, M.; Suriñach, J. (2000). *Tòpics d'econometria*. Barcelona: Editorial UOC.

Dobson, A. J.; Barnett, A. (2008). *An introduction to generalized linear models*. Cleveland: CRC Press.

Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models* (vol. 124). Cleveland: CRC Press.

Guillen, M. (2014). «Regression with Categorical Dependent Variables». En: E. W. Frees; R. Derrig; G. Meyer (eds.). *Predictive Modeling Applications in Actuarial Science* (vol. I). Cambridge: Cambridge University Press.

Harrell, F. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Berlín: Springer.

Hilbe, J. M. (1994). «Generalized linear models». *The American Statistician* (vol. 48, n.º 3, págs. 255-265).

Kabacoff, R. (2015). *R in action: data analysis and graphics with R*. Nueva York: Manning Publications.

McCullagh, P.; Nelder, J. A. (1989). «Generalized Linear Models». *Monograph on Statistics and Applied Probability* (n.º 37).

Nelder, J. A.; Baker, R. J. (1972). «Generalized linear models». *Encyclopedia of Statistical Sciences*. Nueva Jersey: John Wiley & Sons.

Turner, H. (2008). «Introduction to generalized linear models». *Rapport technique*. Viena: Universidad de Economía de Viena.

Wedderburn, R. W. (1974). «Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method». *Biometrika* (vol. 61, n.º 3, págs. 439-447).