Modelos de elección discreta

Román Salmerón Gómez

Grado en Economía

Introducción

Introducción

Hasta el momento se ha trabajado con variables cualitativas incluyéndolas dentro del grupo de variables independientes pero, ¿puede la variable dependiente ser de naturaleza cualitativa? ¿qué ocurre en tal caso? ¿sigue siendo válido el modelo lineal y su estimación por Mínimos Cuadrados?
En ocasiones, analizamos modelos donde la variable dependiente de interés toma valores discretos:

- Variables dependientes binarias (ej: comprar o no comprar, conceder o no un préstamo, tener o no una enfermedad).
- Variables discretas sin ordenación (ej: tren, autobús...).
- Variables discretas con orden (ej: calificación o rating financiero).

En estos casos, el modelo de regresión lineal puede no ser adecuado ya que, por ejemplo, a) los resultados son difíciles de interpretar (no se puede hablar de cambio continuo) y b) la variable dependiente sólo admite valores discretos y puede que sólo no-negativos.
A continuación analizaremos con algo más de profundidad los problemas que surgen al considerar un modelo de regresión lineal clásico en el que la variable dependiente es cualitativa (más concretamente, binaria) y se plantearán las dos principales alternativas que se tienen en este caso: los modelos logit y probit.

De los ejemplos considerados anteriormente, este capítulo se centra fundamentalmente en el caso en el que la variable dependdiente es una variable binaria (dicotómica). Es decir, supondremos que la variable dependiente Y solo puede tomar dos valores:

$$Y = \begin{cases} 1 & \text{, con probabilidad } p \\ 0 & \text{, con probabilidad } 1-p \end{cases}, \tag{1}$$

donde el valor 1 denota que el individuo ha tomado alguna acción. La variable Y, por tanto, sigue una distribución de Bernoulli:

$$\begin{aligned} \Pr(Y=y) &= p^y(1-p)^{1-y}, \\ E[Y] &= p, \\ Var(Y) &= p(1-p). \end{aligned} \tag{2}$$

En tal caso, se estará interesado en analizar cuál es la probabilidad de que el individuo i, dadas sus características (es decir, valores de las variables independientes, X_i), tome una acción (es decir, $Y_i = 1$).

Modelo Lineal de Probabilidad

El Modelo Lineal de Probabilidad consiste simplemente en considerar un modelo de regresión lineal en el que la variable dependiente es binaria, es decir:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, \quad i = 1, \ldots, n, \tag{3}$$

con $u \sim N(0, \sigma^2)$ e Y_i de la forma dada en (1). En este caso, dados los valores x_2, ..., x_k de las variables independientes, se verifica que:

$$\begin{aligned} E[Y_i|X = x] &= \Pr(Y_i = 1|X = x) \cdot 1 + \Pr(Y_i = 0|X = x) \cdot 0 \\ &= \Pr(Y_i = 1|X = x), \\ E[Y_i|X = x] &= E[\beta_0 + \beta_1 X_{2i} + \cdots + \beta_k X_{ki} + u_i|X = x] \\ &= \beta_1 + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}. \end{aligned}$$

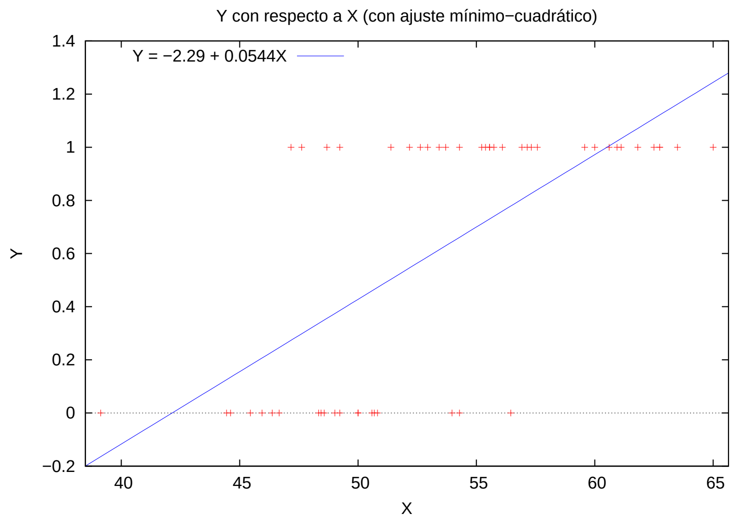
Es decir, la parte derecha de la ecuación (3) debe ser interpretada como la probabilidad de que la variable dependiente sea igual a la unidad:

$$p_i = \Pr(Y_i = 1|X = x) = \beta_1 + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}. \tag{4}$$

Por tanto, β_i es la variación de la probabilidad de que $Y_i = 1$ asociada con una variación unitaria en X_i, manteniendo constantes las otras variables explicativas (con $i = 1, \ldots, k$).

Todo lo que conocido sobre el modelo de regresión lineal se puede aplicar directamente: estimación, contraste de hipótesis, interpretación de los parámetros, etc. Solo debemos recordar que la esperanza condicional es, en este caso, una probabilidad, por lo que $0 \leq E[Y_i|X = x] \leq 1$.

La distribución de la muestra en este tipo de modelos se caracteriza por una nube de puntos de tal manera que las observaciones muestrales se dividen en dos subgrupos: uno formado por las observaciones en las que ocurrió el acontecimiento objeto de estudio ($Y_i = 1$), y otro, por los puntos muestrales en los que no ocurrió ($Y_i = 0$).



Por tanto, el coeficiente de determinación R^2 no es particularmente útil porque no es posible que todos los datos se encuentren exactamente en la recta de regresión ($R^2 = 1$).

Inconvenientes del Modelo Lineal de Probabilidad

- La variable dependiente solo toma valores 0 ó 1, luego los residuos siguen el patrón de una Bernouilli y el supuesto de normalidad de las perturbaciones puede no cumplirse:

	$e_i = Y_i - \hat{Y}_i$	Probabilidad
$Y_i = 1$	$1 - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \cdots - \hat{\beta}_k X_{ki}$	p_i
$Y_i = 0$	$-\hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \cdots - \hat{\beta}_k X_{ki}$	$1 - p_i$

  Este hecho invalidaría el uso de los estadísticos utilizados para contrastar hipótesis ya que estos estadísticos se basan en la hipótesis de normalidad.
- Perturbaciones heteroscedásticas (incumplimiento de la hipótesis de homocedasticidad) ya que su varianza depende de las variables independientes:

$$
\begin{aligned}
Var(e_i) &= E[e_i - E[e_i]]^2 = E[e_i]^2 \\
&= (1 - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \cdots - \hat{\beta}_k X_{ki})^2 p_i \\
&\quad +(-\hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \cdots - \hat{\beta}_k X_{ki})^2 (1 - p_i) \\
&= (1 - p_i)^2 p_i + p_i^2 (1 - p_i) = p_i(1 - p_i).
\end{aligned}
$$

Inconvenientes del Modelo Lineal de Probabilidad

- Las predicciones de la variable dependiente pueden estar fuera del rango [0, 1].
- El modelo lineal de probabilidad implica que el efecto marginal de cada una de las variables explicativas es constante. Este supuesto no es muy razonable ya que es esperable que las variaciones en la probabilidad sean distintos en los valores centrales de las variables dependientes a las producidas en sus extremos.

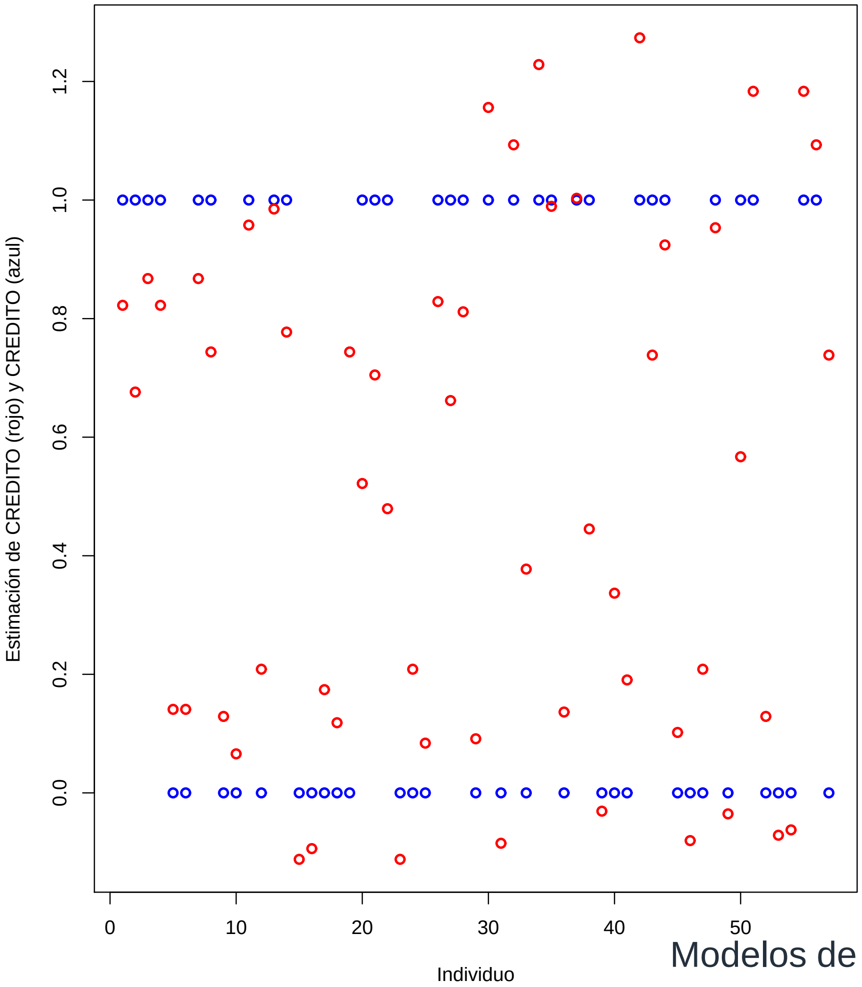
Supongamos que deseamos analizar la influencia que tiene sobre la devolución de un crédito (1 si el crédito es devuelto), $\mathbf{C}$, los ingresos anuales brutos del cliente (en decenas de miles de euros), $\mathbf{I}$, la situación labotal del cliente (0 en paro, 1 contrato temporal, 2 contrato fijo), $\mathbf{L}_0$, $\mathbf{L}_1$ y $\mathbf{L}_2$, y las cargas familiares (1 si el cliente tiene cargas familiares), $\mathbf{F}$.

Planteando una regresión lineal múltiple y estimándola por MCO:

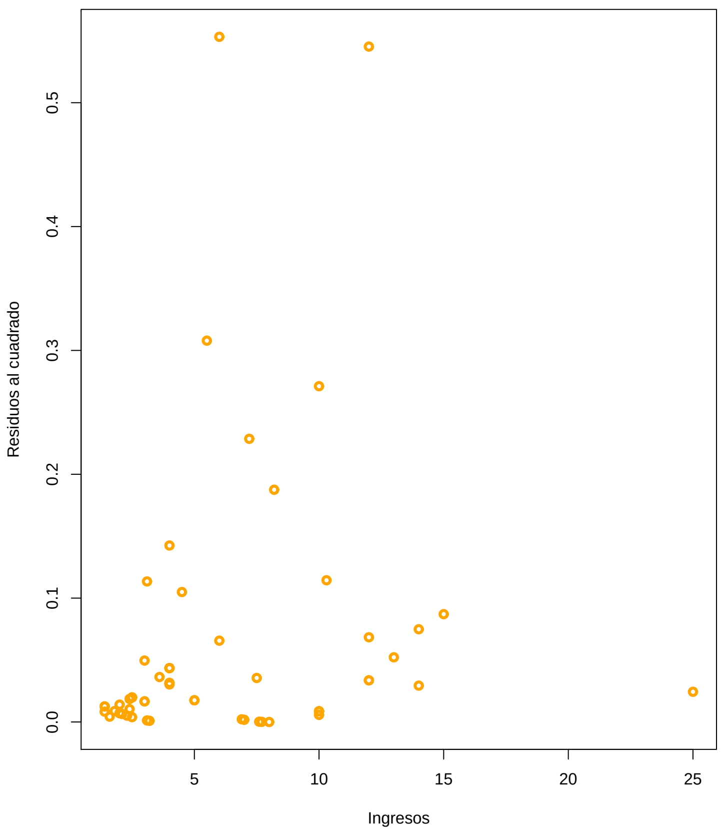
$$C_i = -0.006364 + 0.045123 \cdot I_i + 0.20331 \cdot L_{1i} + 0.648285 \cdot L_{2i} - 0.168883 \cdot F_i.$$

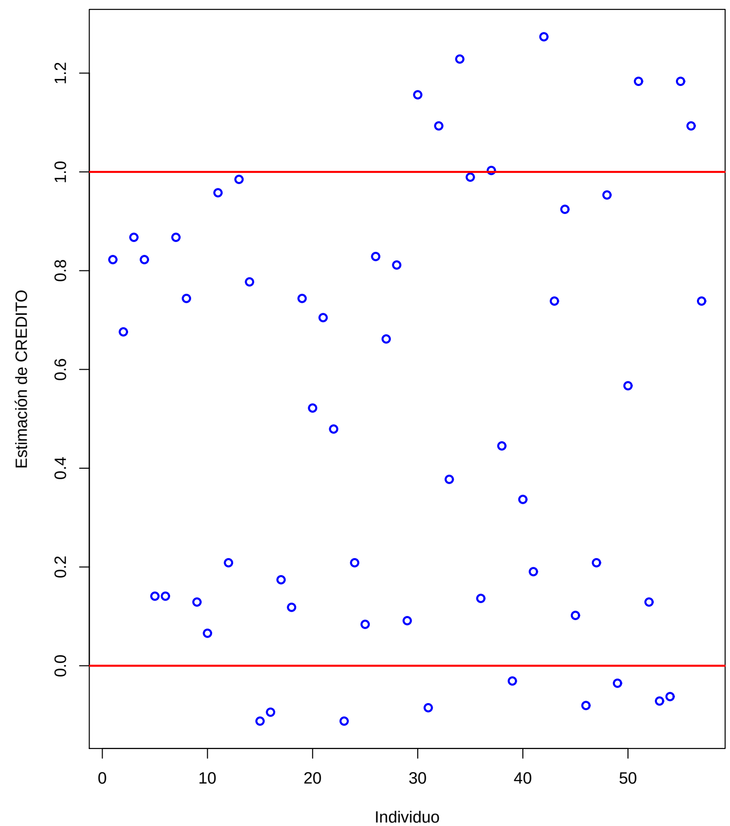
¿Cómo se interpreta la estimación del término constante? ¿Tiene sentido que el efecto marginal del coeficiente de los ingresos sobre la devolución del crédito sea constante? Es decir, ¿la probabilidad de devolver el crédito es indiferente al nivel de ingresos?

Representando la variable dependiente y su estimación, ¿tiene sentido el coeficiente de determinación? ¿Captará algo?



Representando los residuos en función de los ingresos se observa una tendencia creciente, lo cual nos hace pensar en la posible existencia de heterocedasticidad. También se obtienen estimaciones mayores que uno y menores que cero.





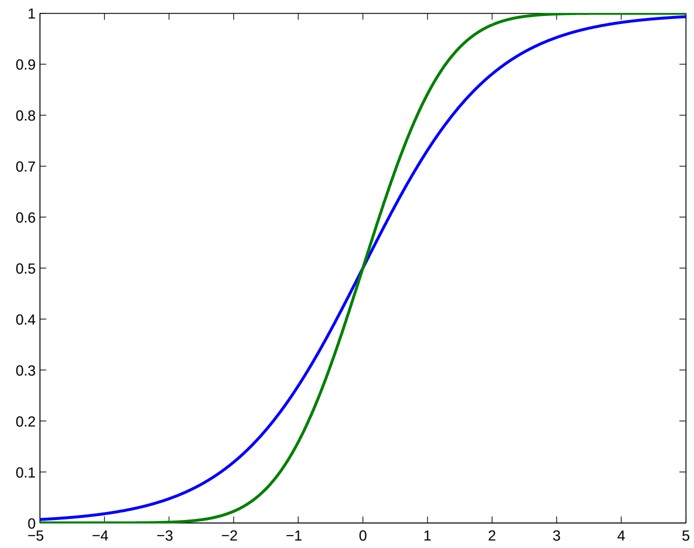
Los modelos Logit y Probit

Los modelos Logit y Probit

Como se ha puesto de relevancia, el modelo lineal de probabilidad presenta importantes inconvenientes que desaconsejan su uso ante variables dependientes binarias y crea la necesidad de recurrir a otros tipos de modelos.
Las regresiones Probit y Logit son modelos de regresión no lineales diseñados específicamente para variables dependientes binarias. Se trata de adoptar una formulación no lineal que obligue a que los valores estimados estén entre 0 y 1 ya que, como hemos visto, la regresión con una variable binaria dependiente Y modeliza la probabilidad de que $Y = 1$.
La regresión Logit utiliza una función de distribución logística, mientras que la regresión Probit utiliza una función de distribución normal estándar. Ambas funciones de distribución de probabilidad dan lugar a probabilidades ente 0 y 1, y presentan un crecimiento no lineal (con mayores incrementos en la parte central). De esta forma se resuelven dos de los problemas anteriormente señalados.

Los modelos logit y probit comparten practicamente las mismas carecterísticas: son modelos no lineales que son estimados por los métodos de mínimos cuadrados no lineales o máxima verosimilitud, donde la interpretación de los coeficientes no es tan inmediata como en el modelo lineal de probabilidad. Además, en ambos casos hay que buscar una medida alternativa al coeficiente de determinación para medir la bondad del ajuste realizado.

La única diferencia entre ambos modelos es que la función logística (curva azul) tiene colas más anchas, por lo que la probabilidad de éxito será mayor en los extremos cuando se use el modelo logit.



Modelo Logit

El modelo de regresión Logit se basa en la función logística:

$$f(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+\frac{1}{e^z}} = \frac{e^z}{1+e^z},$$

la cual está acotada entre 0 y 1 ya que:

$$\lim_{z\to-\infty} f(z) = 0, \quad \lim_{z\to\infty} f(z) = 1,$$

y, como se muestra en la Figura 17, presenta una forma de S que se ajusta al crecimiento no lineal deseado (leves incrementos en los extremos y mayores en la parte central).
El modelo de regresión Logit será de la forma:

$$Y_i = f(Z_i) + u_i, 1 = 1, \ldots, n, \tag{5}$$

donde $Z_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$.

Dados los valores de las variables independientes $x_2, ..., x_k$, las probabilidades de que la variable dependiente tome los valores 1 y 0 son:

$$\begin{aligned}\Pr(Y = 1|x_2, ..., x_k) &= E(Y_i|X = x) = \frac{e^{z_i}}{1 + e^{z_i}}, \\ \Pr(Y = 0|x_2, ..., x_k) &= 1 - \frac{e^{z_i}}{1 + e^{z_i}} = \frac{1}{1 + e^{z_i}},\end{aligned}$$

con $z_i = \beta_1 + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$.

Si se estima el modelo sobre el crédito planteando un modelo Logit se obtiene que:

$$\hat{z}_i = -20.9772 + 0.4847 \cdot I_i + 18.2687 \cdot L_{1i} + 21.9597 \cdot L_{2i} - 2.1951 \cdot F_i.$$

Ahora, por ejemplo, se tiene garantizado obtener estimaciones en el intervalo [0, 1]: la mayor es 0.9995771 y la menor 1.702466e-10.

Para el individuo medio, $x = (1, 6.1701754, 0.3684211, 0.3684211, 0.4561404)$, se tiene que $\hat{z}_i = -4.166983$ y, por tanto, una probabilidad de éxito del 1.52624%. ¿Tiene sentido el cálculo realizado?

Para un individuo (individuo 1) con ingresos medios que tenga contrato temporal y no tenga cargas familiares, $x = (1, 6.170175, 1, 0, 0)$, la probabilidad de devolver el crédito para este individuo es del 57.004% ($\hat{z}_i = 0.2820397$).

Si se considerase que sí tiene cargas familiares (individuo 2) la probabilidad pasaría a ser del 12.86% ($\hat{z}_i = -1.913109$).

El modelo de regresión Probit se basa en la distribución de probabilidad acumulada de una normal tipificada:

$$\Phi(z) = \Pr(Z \geq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{s^2}{2}} ds,$$

donde $Z \sim N(0,1)$ y es tal que, dados los valores $x_2, ..., x_k$ de las variables independientes, se verifica que:

$$\Pr(Y = 1 | x_2, ..., x_k) = \Phi(z_i),$$

con $z_i = \beta_1 + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$ tal que:

$$Y = \begin{cases} 1 & \text{si } z_i > 0 \\ 0 & \text{si } z_i < 0 \end{cases} .$$

Ejemplo

Si se estima el modelo sobre el crédito planteando un modelo Probit se obtiene que:

$$\hat{z}_i = -6.4991 + 0.2562 \cdot I_i + 4.9853 \cdot L_{1i} + 7.1625 \cdot L_{2i} - 1.2033 \cdot F_i.$$

Igualmente se tiene garantizado obtener estimaciones en el intervalo [0, 1].
En este caso, para el individuo 1, $x = (1, 6.170175, 1, 0, 0)$, la probabilidad de devolver el crédito para este individuo es del 52.675% ($\hat{z}_i = 0.06711711$), mientras que para el individuo 2 la probabilidad pasaría a ser del 12.79% ($\hat{z}_i = -1.1362$).
Se observa que las probabilidades de éxito son muy parecidas a las obtenidas por el modelo Logit.

Efecto marginal

En los modelos lineales (como el modelo lineal de probabilidad) la derivada parcial de la variable dependiente, Y, con respecto a cada una de las variables explicativas, X_j, $j = 1, \ldots, p$, es la constante β_j, y se interpreta como el cambio producido en Y cuando X_j aumenta una unidad. Puesto que los modelos logit y probit son no lineales, esta interpretación no es correcta.

En el modelo Logit, partiendo de (5), la derivada parcial anterior es:

$$\frac{\partial Y_i}{\partial X_{ji}} = \frac{e^{-Z_i}}{(1 + e^{-Z_i})^2} \cdot \beta_j, \quad j = 1, \ldots, k,$$

mientras que en el probit:

$$\frac{\partial Y_i}{\partial X_{ji}} = \phi(z_i)\beta_j, \quad j = 1, \ldots, k,$$

siendo ϕ la función de densidad de la distrbución normal tipificada.

Por tanto, el efecto marginal en ambos modelos depende de los valores que toman las variables explicativas (ya no es constante: uno de los objetivos perseguidos por estos modelos).

Pueden, por tanto, calcularse los efectos marginales para cada observación de la muestra. Puesto que esto supone una limitación importante, alternativamente, los efectos marginales suelen evaluarse para el valor medio de las variables explicativas. Si bien, hay que tener espacial cuidado cuando dentro de éstas hay variables no cuantitativas.

Puesto que la exponencial y la función de densidad ϕ son siempre positivas, queda claro que el signo de los coeficientes indica la dirección del efecto marginal. Es decir, un signo positivo indicará una relación directa (aumento de la probabilidad de que Y se igual a 1), mientras que uno negativo inversa (disminución de la probabilidad de que Y se igual a 1).

Odd ratio

Debido a las dificultades existentes para interpretar el efecto marginal, en la práctica se calcula la razón entre las probabilidades p_i y $1 - p_i$, cociente denominado *odd*, es decir:

$$\frac{p_i}{1 - p_i}.$$

El *odd* se interpreta como el número de veces que es más probable que ocurra el fenómeno o suceso frente a que no ocurra dada una observación concreta. Por tanto, el *odd* se asocia a individuos cocnretos y no a las variables independientes. En el caso del modelo Logit, el *odd* se puede calcular fácilmente a partir de:

$$\frac{p_i}{1 - p_i} = \frac{\frac{e^{z_i}}{1+e^{z_i}}}{\frac{1}{1+e^{z_i}}} = e^{z_i},$$

donde $z_i = \beta_1 + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$.

Por otro lado, el *odd-ratio* es un cociente de *odds* y hace referencia a un grupo determinado, es decir, hace referencia a un cociente entre individuos que han experimentado el suceso y los que no, asociado a las variaciones de una variable independiente concreta (la cual determina el grupo de interés).

En este caso, dados dos individuos diferentes h y l, el *odd-ratio* responde a la siguiente expresión:

$$\frac{\frac{p_h}{1-p_h}}{\frac{p_l}{1-p_l}}.$$

El *odd-ratio* indica cuanto es más probable que se produzca el suceso de interés, $Y = 1$, para la observación h o para la observación l de la variable independiente usada.

Por tanto, el *odd-ratio* representa el cambio producido en el odds ante un cambio en una variable explicativa.

Odd ratio

Una vez más, en el caso del modelo Logit, se puede calcular de manera sencilla el *odd-ratio*. Así, el *odd-ratio* asociado a un cambio de x_{jh} a x_{jl}, $h \neq l$, $h, l = 1, \ldots, n$ en la variable X_j, $j = 1, \ldots, k$, supuesto que el resto de variables permanecen constantes, viene dado por:

$$\frac{e^{z_h}}{e^{z_l}} = e^{\beta_j (x_{jh} - x_{jl})}.$$

Por tanto, si la diferencia $x_{jh} - x_{jl}$ es igual a s, el *odd-ratio* representa el cambio que se produce en el odds cuando la variable dependiente X_j incrementa s unidades. Es decir, el *odd-ratio* ya no depende de las observaciones, sino solamente de la variable X_j.

En tal caso:

- Si no existe relación entre la variable dependiente y la variable en estudio, el *odd-ratio* toma el valor uno.
- Si la variable dependiente incrementa la probabilidad de la explicada, el *odd-ratio* será superior a uno tanto mayor cuanto más elevada sea esta relación.
- Si la variable dependiente disminuye la probabilidad de la explicada, el *odd-ratio* será menor que uno.

Puesto que el efecto marginal tiene el mismos signo de la estimación de los coeficientes, se tiene que para signos estimados positivos la probabilidad de que se devuelva el crédito aumenta. Y para los negativos que disminuye.

En el caso del modelo del crédito planteado, tanto para el modelo Logit como el Probit, el único coeficiente significativamente distinto de cero es el referente a los ingresos. Puesto que tiene signo positivo, se tiene que un aumento en los ingresos supone un aumento en la probabilidad de devolver el crédito.

Se podría calcular el efecto marginal para un individuo concreto y así poder cuantificar el aumento correspondiente en cada caso. ¿Cuál sería el efecto marginal de los individuos considerados anteriormente?

Por otro lado, en el Logit, los *odds* asociados a los individuos anteriores son 1.325831 y 0.1476207, respectivamente. Para el primer individuo, se obtiene que es 1.3258 veces más probable que devuelva el crédito frente a que no lo haga, mientras que para el segundo, es 6.774 veces menos (¡ojo!) probable que devuelva el crédito frente a que no lo haga.

En el Probit se obtienen valores muy similares: 1.113 y 6.816.

Si se hace el cociente de dos *odds*, se obtiene el número de veces más probable que es devolver un crédito frente a no hacerlo asociado a los cambios producidos en las variables explicativas.

Así, con el cociente de los *odds* calculados anteriormente se obtiene que en un individuo con ingresos medios, con contrato temporal y con cargas (cargas=1) es 8.9813 veces menos probable que se devuelva el crédito frente a no hacerlo que cuando no tiene cargas (cargas=0). Adviértase que al obtener un *odd-ratio* inferior a 1 se ha cambiado la interpretación de más probable por menos probable.

El *odd-ratio* anterior se asocia a incrementos de una unidad en la variable referente a la carga familiar. Si se quieren ver incrementos de más unidades, por ejemplo en el Logit, hay que multiplicar el coeficiente estimado por dicha unidad y posteriormente calcular la exponencial. Así, mediante $e^{10 \cdot 0.4847} = 127.313$, se tiene que un individuo que tiene unos ingresos en 10000 euros superiores a otro (aumento de 10 unidades en la variable ingresos) es 127.313 veces más probable que devuelva el crédito frente a que no lo haga suponiendo que el resto de variables permanecen constantes.

Coeficiente de McFadden

En los modelos Logit y Probit, debido a que son modelos no lineales, no podemos utilizar el coeficiente de determinación clásico para medir la bondad del ajuste. Recordemos que este era uno de los problemas que surgían en el modelo lineal de probabilidad. En su lugar, se utiliza el pseudo R^2 de McFadden:

$$\tilde{R}^2 = 1 - \frac{\ln L}{\ln L_r},$$

donde $\ln L$ es el logaritmo neperiano de la función de verosimilitud del modelo sin restricciones (el modelo con todas las variables explicativas) y $\ln L_r$ es el logaritmo neperiano de la función de verosimilitud del modelo restringido (solo incluye el término independiente del modelo).
Valores próximos a cero indicará que el modelo no es adecuado al contrario que valores próximos a uno.

Otra opción para analizar la bondad del modelo es contabilizar el porcentaje de aciertos del modelo teniendo en cuenta que, por ejemplo, las probabilidades predichas por encima de 0.5 contabilizan como $Y_i = 1$ y menores que 0.5 estiman $Y_i = 0$:

	$\hat{Y}_i = 1$	$\hat{Y}_i = 0$
$Y_i = 1$	A	B
$Y_i = 0$	C	D

En los casos A y D se habra predicho correctamente el valor de Y, por tanto, la proporción de aciertos vendrá dada por el cociente $\frac{A+D}{n}$.
Si se desea ser exigente con el modelo (lo recomendado), en lugar de usar el umbral del 0.5 se debe usar la proporción de éxitos (de unos) que hay en la variable dependiente.

En el caso de la devolución de crédito, considerando como umbral la proporción de unos en la variable dependiente (0.5087719), se tiene la siguiente tabla de clasificación de aciertos:

	$\hat{Y}_i = 1$	$\hat{Y}_i = 0$	Porcentaje aciertos
$Y_i = 1$	27	2	93.103%
$Y_i = 0$	2	26	92.857%
			92.982%

Es decir, se han clasificado correctamente el 92.982% de los individuos (53 de los 57).

Para realizar contrastes de significación individual sobre los coeficientes:

$$\left.\begin{array}{l} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{array}\right\},$$

nos basaremos en que los estimadores siguen una distribución normal $\hat{\beta}_j \sim N(\beta_j, \text{Var}(\beta_j))$. Por tanto, para tomar una decisión en el contraste utilizamos la siguiente regla de decisión:

$$\text{Se rechaza } H_0 \text{ si } \left| \frac{\hat{\beta}_j}{\sqrt{\text{Var}(\hat{\beta}_j)}} \right| \geq Z_{1-\alpha/2},$$

donde $P[Z < Z_{1-\alpha/2}] = 1 - \alpha/2$ con $Z \sim N(0,1)$.

Adviértase que la obtención de la matriz de varianzas-covarianzas de los coeficientes, β_j, no es una tarea fácil. Por suerte todos los programas informáticos lo realizan automáticamente, por lo que se puede realizar inferencia en la forma habitual.

Significación conjunta de los coeficientes

Para realizar contrastes de significación conjunta sobre todos los coeficientes (o un subconjunto), se puede utilizar el contraste de la razón de verosimilitudes:

$$\left.\begin{array}{l} H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0 \\ H_1 : \text{ en caso contrario} \end{array}\right\} .$$

El estadístico de contraste es:

$$-2 \ln \frac{L(\hat{\beta}_r)}{L(\hat{\beta})} \sim \chi^2_q,$$

donde $L(\hat{\beta}_r)$ es la verosimilitud del modelo restringido, es decir, del modelo en el que se impone la H_0, $L(\hat{\beta})$ es la verosimilitud del modelo sin restricciones y q es el número de restricciones.