Modelización predictiva: introducción a los modelos lineales generalizados

PID_00247912

Montserrat Guillén Estany

Tiempo mínimo de dedicación recomendado: 2 horas

Universitat Oberta de Catalunya

Índice

Introducción

Los modelos de regresión permiten especificar una relación causa-efecto entre variables. Al utilizar datos para estimar dichos modelos se obtienen las estimaciones que permiten efectuar predicciones de los efectos a partir de escenarios posibles. 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, 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 la *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 a través de 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.

Los modelos que se tratan es este módulo son modelos más avanzados. Una de las propiedades 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 forma tan específica, ya sea por su propia naturaleza o porque ya han sido registrados como variables no continuas. Un ejemplo de una variable no continua es tomar una decisión

entre conceder un crédito a un cliente o no, que al ser una decisión dicotómica entre únicamente dos posibilidades, no puede tomar más que esos dos valores. En el caso del ejemplo de datos del módulo «Introducción a la estadística», el nivel salarial estaba medido de forma discreta; en este caso los posibles valores de la variable son 1, 2, 3 o 4 y, dan información sobre el nivel y no sobre el sueldo exacto.

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

ID	Nombre	Edad	Sexo	Estado civil	Número de hijos	Nivel salarial	Crédito solicitado	Préstamo hipotecario	Motivo de préstamo
567	José Pérez	46	H	C	2	3	20.000	65.000	Vehículo
765	María Sol	34	M	S	0	2	5.000	0	Estudios
965	Simón Martín	52	H	C	2	4	350.000	0	Reformas

Fuente: Ejemplo del capítulo "Introduccion a la estadística"

Los elementos clave de un modelo de regresión lineal múltiple clásico son los siguientes:

- La **variable dependiente** es de naturaleza continua, y condicionada a los valores de las explicativas, tiene un comportamiento aleatorio normal (en forma de campana de Gauss).

- Las **variables explicativas** son conocidas de antemano y constituyen los factores causales de la respuesta, que se obtendrá como una combinación lineal de dichas variables explicativas.

- El término de 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.

Es importante recordar que en los modelos predictivos más avanzados que se tratarán en este módulo, 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. ¿Qué son los modelos lineales generalizados?

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 publica junto a 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 modelos lineales generalizados

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…
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…
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…)	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 (ver por ejemplo Wedderburn, 1974). 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 aproximación sigue siendo las más ampliamente utilizada.

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.

La componente denominada «**predictor lineal**» de este modelo es:

$$\beta_0 + \beta_1 Edad_i + \beta_2 Mujer_i + \beta_3 SL1_i + \beta_4 SL2_i + \beta_5 Ingresos_i$$

Es importante destacar que:

- En este ejemplo tenemos seis parámetros (los parámetros beta) 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. Por ello 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 se hayan estimados 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.

La **componente aleatoria** 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. Se establece el siguiente comportamiento estocástico (variable Bernoulli):

$$Y_i = \begin{cases} 1, \text{ con probabilidad } p_i \\ 0, \text{ con probabilidad } 1\text{-}p_i \end{cases}$$

Respecto a esta componente hay que destacar que la probabilidad de comprar o no es individual de cada persona y 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 p_i. Es decir, la esperanza matemática $E(Y_i)$ se calcula como:

$$E(Y_i) = 1 \times p_i + 0 \times (1 - p_i) = p_i$$

En definitiva, el valor que va a predecirse es la probabilidad de que la variable dependiente tome el valor 1, es decir $E(Y_i) = p_i = Prob(Y_i = 1)$.

La tercera componente es una **función de ligadura** que hace corresponder cada valor de p_i a un valor del predictor lineal y que cumple las condiciones exigibles en el modelo lineal generalizado. Es decir, debe ser una función monótona, continua y diferenciable.

La elección de la función de ligadura determinará finalmente el tipo de modelo lineal generalizado que se haya establecido.

$$E(Y_i) = p_i = f(\beta_0 + \beta_1 Edad_i + \beta_2 Mujer_i + \beta_3 SL1_i + \beta_4 SL2_i + \beta_5 Salario_i)$$

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

$$Prob(Y_i = 1) = \frac{\exp(\beta_0 + \beta_1 Edad_i + \beta_2 Mujer_i + \beta_3 SL1_i + \beta_4 SL2_i + \beta_5 Ingresos_i)}{1 + \exp(\beta_0 + \beta_1 Edad_i + \beta_2 Mujer_i + \beta_3 SL1_i + \beta_4 SL2_i + \beta_5 Ingresos_i)}.$$

Todos los modelos lineales generalizados obedecen a las tres partes consideradas y por lo tanto, se determinan fijando:

1) ¿Cuáles son las variables que participan en la determinación del predictor lineal?

2) ¿Qué distribución estadística se utiliza para la variable dependiente?

3) ¿Cuál es la función de ligadura?

2. Estimación por máxima verosimilitud

Para proceder a la estimación de los parámetros es necesario utilizar un método que permita optimizar un criterio y que conduzca a los mejores valores posibles para las «betas». La función de verosimilitud se define a partir de los datos observados, del predictor lineal y la función de ligadura, así como del comportamiento estocástico de la variable dependiente. La función de verosimilitud 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 las mismas. 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 sobre 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 dimensión seis, cuyo dominio sería el producto de seis números reales y su recorrido el intervalo [0,1], puesto 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. Modelo de regresión logística o modelo logit

El modelo de regresión logística especifica que:

$$Prob\,(Y_i = 1) = \frac{exp\,(\beta_0 + \beta_1 X1_i + \beta_2 X2_i + \ldots + \beta_k Xk_i)}{1 + exp\,(\beta_0 + \beta_1 X1_i + \beta_2 X2_i + \ldots + \beta_k Xk_i)}$$

Donde *X*1, ..., *Xk* se refieren a las variables explicativas introducidas en el modelo.

Mientras que la inversa de dicha relación, lo que normalmente se denomina la función de ligadura (*link function*), es:

$$\beta_0 + \beta_1 X1_i + \beta_2 X2_i + \ldots + \beta_k Xk_i = logit\,(p_i) = ln\,[\frac{p_i}{(1 - p_i)}]$$

Denominamos a $\frac{p_i}{(1-p_i)}$ *odds* o cociente de probabilidades, y a su logaritmo *log-odds*. Por lo tanto, en el modelo de regresión logística el predictor lineal es igual al logaritmo del cociente de probabilidades. Como lo que se pretende predecir es la probabilidad de ocurrencia de algún suceso, entonces el modelo logit es un modelo que permite predecir la probabilidad de dicha ocurrencia.

El modelo de regresión logística se utiliza en muchos ámbitos; 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; etc. Lo mejor del modelo logit es que la predicción siempre se encuentra entre 0 y 1 por lo tanto siempre es interpretable como una probabilidad, o como un valor en porcentaje. La función logit tiene una forma de «S» de forma que si el predictor crece se acerca a 1 y si decrece se acerca a 0. Por lo tanto, cuanto mayor es el predictor más alta es la probabilidad, y cuando el predictor es muy negativo, la probabilidad se acerca a 0.

4. Interpretación de coeficientes y *odds-ratios*

Los parámetros se interpretan en el modelo de regresión logística binaria de una manera más complicada, en comparación con cómo se interpretan en el contexto del modelo de regresión lineal.

Se necesita un poco de práctica para acostumbrarse a hablar de probabilidades ajustadas en lugar de valores ajustados. Al interpretar los resultados de la estimación de un modelo de regresión logística, 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, entonces cuando el factor predictivo (variable independiente) aumenta, la probabilidad del suceso modelado disminuye. Por ejemplo, en el caso del ejemplo descrito anteriormente, si el parámetro asociado al salario fuese positivo, se interpretaría que a mayor salario, mayor es la probabilidad de que la persona se decida por comprar la fragancia ofertada por el fabricante. Por otro lado, si el parámetro asociado a la edad fuese negativo, entonces significaría que cuanto mayor es la persona, menos probable es que tenga intención de comprar el producto.

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 de la probabilidad del suceso cuando todos los regresores son iguales a cero.

El cociente de los *odds* (llamado *odds-ratio*) es el nombre dado a la exponencial de un parámetro. Supongamos que el individuo i cambia su *k*-ésimo predictor, *Xk*, en una unidad. Por ejemplo, supongamos que inicialmente esa característica *Xk* es igual a *c* y cambia entonces para ser igual a $c + 1$. Entonces se ve fácilmente que:

$$\exp(\beta_j) = \frac{Prob\,(Y_i = 1 | Xj_i = c+1) \,/\, Prob\,(Y_i = 0 | Xj_i = c+1)}{Prob\,(Y_i = 1 | Xj_i = c) \,/\, Prob\,(Y_i = 0 | Xj_i = c)} = \frac{e^{\beta_j (c+1)}}{e^{\beta_j c}}$$

Luego,

$$\beta_j = \ln\left(\frac{Prob\,(Y_i = 1 | Xj_i = c+1)}{Prob\,(Y_i = 0 | Xj_i = c+1)}\right) - ln\left(\frac{Prob\,(Y_i = 1 | Xj_i = c)}{Prob\,(Y_i = 0 | Xj_i = c)}\right).$$

Si la *j*-ésima variable explicativa es continua, también podemos deducir que el parámetro β_j es el cambio proporcional en la *odds-ratio* o en la elasticidad económica. Un concepto conectado al *odds-ratio* es la noción de razón de riesgo. La razón de riesgo es la proporción de probabilidades de sucesos cuando el indicador de predicción cambia en una unidad. Esto es especialmente importante para ver el cambio relativo en la probabilidad de un «suceso» cuando un indicador de riesgo está presente en lugar de ausente. La definición es:

$$RR_j = \frac{Prob\,(Y_i = 1 | Xj_i = c + 1)}{Prob\,(Y_i = 1 | Xj_i = c)}$$

Las relaciones de riesgo dependen del vector de variables explicativas y no pueden expresarse como una función de un solo parámetro.

Supongamos que en el ejemplo de compra del perfume obtenemos una estimación para el parámetro de la edad igual a -0,03. Al ser negativo ya podemos afirmar que a mayor edad, menor probabilidad de comprar el producto, siendo el resto de características las mismas. El *odds-ratio* sería igual a exp(-0,03)=0,97. Eso quiere decir que cuando aumenta un año la edad de la persona la probabilidad de comprar se multiplica por 0,97, lo que quiere decir que disminuye un 3 %.

5. Matriz de confusión

Las estimaciones de parámetros se indican con un acento ^ encima de la beta. Una vez obtenidos, podemos predecir la probabilidad del suceso modelado por la respuesta. Dadas unas características podemos calcular la probabilidad ajustada, que se basa en la siguiente expresión:

$$\widehat{Prob}\,(Y_i = 1) = \frac{exp\,(\hat{\beta}_0 + \hat{\beta}_1 X1_i + \hat{\beta}_2 X2_i + \ldots + \hat{\beta}_k Xk_i)}{1 + exp\,(\hat{\beta}_0 + \hat{\beta}_1 X1_i + \hat{\beta}_2 X2_i + \ldots + \hat{\beta}_k Xk_i)}$$

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 dichos 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 nuestro ejemplo, consideremos a una persona que elige comprar la fragancia, pero que tiene unas características totalmente alejadas de los consumidores más propensos. 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 a 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 casos 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 de ser bajo.

El número de resultados denominados falsos positivos corresponde a casos en 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, significan que el modelo predice que el cliente es probable que vaya a comprar, por lo que es posible que se haya hecho una campaña publicitaria para llegar a este colectivo, 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 mien-

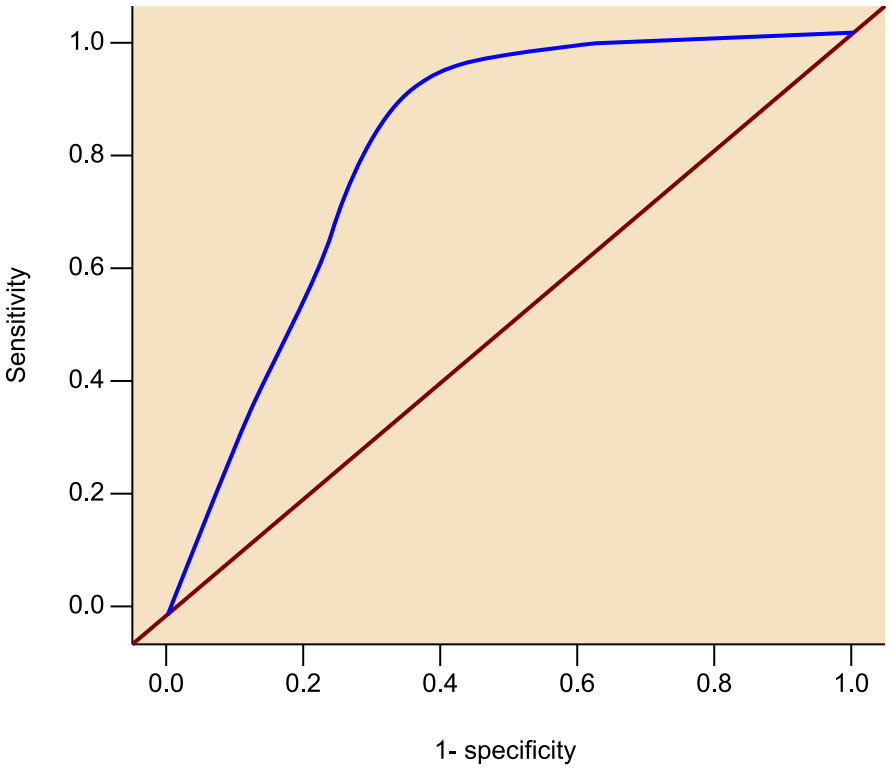
tras que la predicción de la probabilidad de hacerlo es baja, inferior al 50%, corresponden a un total de 10 casos. Este modelo tiene solo unos pocos falsos negativos. Y si se utilizara con fines predictivos, significa que solo habría unos cuantos clientes que, aun no respondiendo al perfil del comprador, 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 %.

6. Curva ROC. Métricas AUROC y KS

Existe una herramienta gráfica para evaluar la bondad de ajuste en la regresión logística. Presentamos la curva *Receiver Operating Characteristic*, también conocida como la curva ROC. 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 una forma de seleccionar modelos posiblemente óptimos y subóptimos basada 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 situada entre la curva y la diagonal, simplemente llamada AUROC (*Area Under the ROC*). El modelo cuya área sea superior es el preferido. Otra forma de medir la calidad del ajuste del modelo es a través de la prueba de Kolmogorov-Smirnov, en la que se compara la distribución de las probabilidades estimadas por el modelo en los dos grupos de respuesta considerados. Si es diferente, se considera que el modelo predictivo discrimina bien.

Figura 1: Ejemplo de curva ROC



7. Selección de umbral y otros modelos

Unos 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.

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 los 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:

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

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 quiere decir 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.

Modelo Poisson

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 características iguales.

Bibliografía

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

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

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

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

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

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

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

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

McCullagh, P.; y 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.

Turner, H. (2008). *Introduction to generalized linear models*. Rapport technique, Vienna University of Economics and Business.

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

Enlace de internet

Curvas ROC: Elección de puntos de corte y área bajo la curva (AUC). Disponible en: <https://www.bioestadistica.uma.es/analisis/roc1/>