Modelització predictiva: introducció als models lineals generalitzats

PID_00247910

Montserrat Guillén Estany

Temps mínim de dedicació recomanat: 2 hores

Universitat Oberta de Catalunya

Índex

Introducció

Els models de regressió permeten especificar una relació causa-efecte entre variables. En utilitzar dades per a estimar aquests models s'obtenen les estimacions que permeten efectuar prediccions dels efectes a partir d'escenaris possibles. Un exemple senzill de regressió lineal simple és el que relaciona el pes d'una persona amb la seva altura. En aquest cas es considera que el pes (variable dependent, variable resposta, explicada o efecte) és conseqüència de l'altura (variable independent, variable tractament, explicativa o causa), a part de molts altres factors. En un model de regressió lineal simple, solament hi ha una variable dependent i una d'independent. S'estableix que la relació és una recta (anomenada *recta de regressió*) i que a partir d'un conjunt de dades, l'estimació de la recta permet efectuar prediccions. La predicció del pes d'una persona en funció de la seva altura s'obté mitjançant la recta estimada i fixant que aquest pes correspon al valor que pren la recta en el punt concret de l'altura. En aquest exemple, queda patent que la predicció és un «valor esperat», que s'obté de l'ajustament a un model molt simple que han proporcionat unes dades històriques.

El pas de la regressió lineal simple a la regressió lineal múltiple sorgeix de la necessitat d'incorporar més factors en l'explicació de la variable dependent. En l'exemple de pes i altura, es poden tenir en compte característiques com la quantitat mitjana de calories diàries ingerides, el sexe, o si es fa algun tipus d'activitat física habitualment. Per tant, amb un model de regressió lineal múltiple, en considerar-se més factors, es pot fer una predicció més precisa que en un model simple si es coneix la informació sobre aquestes característiques. En el model múltiple hi ha una variable dependent i diverses d'independents.

Els models que es tracten en aquest mòdul són models més avançats. Una de les propietats que tenen els models de regressió lineal simple o múltiple és precisament la linealitat. Aquesta hipòtesi pot ser excessivament restrictiva quan la realitat no ho és. Per això, els models no lineals estableixen relacions causa-efecte més complexes, que no solament comporten un repte per a elegir la forma de la relació (parabòlica, exponencial, multinomial, etc.), sinó que alhoras compliquen l'estimació del model.

La segona de les restriccions importants dels models de regressió lineal clàssics és considerar que la variable dependent té un comportament continu; és a dir, pren valors amb decimals. Aquest és el cas del pes, que podria arribar a mesurar-se amb molta precisió. No obstant això, en la pràctica hi ha nombrosos fenòmens que no poden mesurar-se de manera tan específica, sia per la seva pròpia naturalesa o perquè ja han estat registrats com a variables no contínues. Un exemple d'una variable no contínua és prendre una decisió entre concedir un crèdit a un client i no fer-ho, que en ser una decisió dicotòmica entre úni-

cament dues possibilitats no pot prendre més que aquests dos valors. En el cas de l'exemple de dades del mòdul «Introducció a l'estadística», el nivell salarial estava mesurat de manera discreta; en aquest cas els possibles valors de la variable són 1, 2, 3 o 4, i donen informació sobre el nivell i no sobre el sou exacte.

Taula 1. Dades de sol·licitud de préstecs d'una entitat bancària

ID	Nom	Edat	Sexe	Estat civil	Nombre de fills	Nivell salarial	Crèdit sol·licitat	Préstec hipotecari	Motiu del préstec
567	José Pérez	46	H	C	2	3	20.000	65.000	Vehicle
765	María Sol	34	M	S	0	2	5.000	0	Estudis
965	Simón Martín	52	H	C	2	4	350.000	0	Reformes

Font: Exemple del capítol «Introducció a l'estadística»

Els elements clau d'un model de regressió lineal múltiple clàssic són els següents:

- La **variable dependent** és de naturalesa contínua i, condicionada als valors de les explicatives, té un comportament aleatori normal (en forma de campana de Gauss).

- Les **variables explicatives** són conegudes per endavant i constitueixen els factors causals de la resposta, que s'obtindrà com una combinació lineal d'aquestes variables explicatives.

- El terme d'error recull tots els factors que influeixen en el valor observat de la variable dependent no recollits per les variables explicatives, i se suposa que té una mitjana nul·la (els errors de vegades són positius i de vegades negatius). En la inferència sobre els models de regressió lineal, se suposa que el terme d'error segueix un comportament normal. La predicció que proporciona el model de regressió lineal és un valor esperat de la resposta en funció dels valors de les variables explicatives.

És important recordar que en els models predictius més avançats que es tractaran en aquest mòdul, tal com ja ocorria en els models de regressió lineal clàssics, les variables explicatives poden ser de qualsevol tipus, és a dir: contínues, qualitatives dicotòmiques, politòmiques o discretes. No obstant això, serà la naturalesa de la variable dependent la que determinarà quin model predictiu és necessari utilitzar.

1. Què són els models lineals generalitzats?

Els **models lineals generalitzats** van ser formulats inicialment en 1972 i 1974 per dos professors anglesos, John Nelder i Robert Wedderburn respectivament. No obstant això, posteriorment l'impulsor va ser Peter McCullagh, que en 1983 publica, al costat de John Nelder, el llibre *Generalized Linear Models*. La possibilitat d'utilitzar una creixent potència computacional en una dècada en la qual es va popularitzar l'ús d'ordinadors personals va permetre que els nous models poguessin ser a l'abast d'investigadors de tots els àmbits, des de la medicina i la biologia fins a les ciències socials, l'economia i el màrqueting. Això va permetre una expansió ràpida d'aquests models.

Els models lineals generalitzats aconsegueixen unificar un gran nombre de models estadístics predictius, incloent la regressió lineal, la regressió logística (quan la resposta és dicotòmica) i la regressió de Poisson (quan la resposta és un valor que explica el nombre de vegades que ocorre un fenomen).

Els models lineals generalitzats tenen tres components:

1) El predictor lineal.

2) El comportament aleatori de la variable depenent.

3) Una funció que dona una correspondència entre el predictor lineal i el valor esperat de la variable dependent, que es denomina *link* o funció de lligadura.

Solament pot considerar-se que un model lineal generalitzat és ben definit si:

- el predictor lineal és una combinació lineal entre paràmetres i variables explicatives,
- el comportament estocàstic de la variable dependent és dins d'un conjunt concret de distribucions denominat *família exponencial*,
- la funció de lligadura té certes propietats (com ser contínua, monòtona i derivable).

Taula 2. Exemples de models lineals generalitzats

Variable dependent	Variables explicatives	Naturalesa de la variable dependent	Distribució	Models lineals generalitzats possibles
El consumidor decideix comprar o no comprar un producte	Sexe, edat, nivell d'estudis, zona de residència, etc.	Dicotòmica (pren dos valors, SÍ o NO)	Bernoulli	Model de regressió logística, model Probit…
Atorgar un crèdit o no	Edat, nivell salarial, estat civil, nombre de fills	Dicotòmica (pren dos valors, SÍ o NO)	Bernoulli	Model de regressió logística, model Probit…
Nombre de dies que es pernocta en una ciutat durant un viatge vacacional	Edat, nacionalitat, nivell socioeconòmic	Discreta (els valors possibles són 0, 1, 2, 3…)	Poisson	Model de Poisson

Els impulsors dels models lineals generalitzats van proposar un mètode de mínims quadrats reponderat iteratiu per a estimar la màxima versemblança dels paràmetres del model (vegeu, per exemple, Wedderburn, 1974). S'obtenen les estimacions i els seus errors per a obtenir els respectius intervals de confiança amb un nombre petit d'iteracions, i encara que hi ha altres vies d'estimació possibles aquesta aproximació continua essent la més àmpliament utilitzada.

Exemple de model lineal generalitzat

El fabricant de perfums Mi Aroma Favorito vol conèixer com són els consumidors que tenen major propensió a comprar una fragància que llançarà pròximament. Per a això tria un conjunt de 150 persones a les quals pregunta si voldrien comprar-la o no. La resposta és dicotòmica. Els factors que influeixen en aquesta decisió i que seran considerats són *edat*, *sexe*, *situació laboral* i *ingressos mensuals nets*. Vegem en primer lloc que, si bé la variable dependent és dicotòmica, les variables explicatives tenen una naturalesa diferent. L'*edat* és una variable quantitativa discreta. El *sexe* és qualitativa (per tant, es codifica, per exemple, prenent el valor 1 si és una dona i 0 si és un home). La *situació laboral* considera tres possibilitats: treballador per compte propi o d'altri, aturat o altres (jubilat, estudiant, etc.). I finalment la variable *ingressos nets mensuals* es considera contínua. Per a la *situació laboral* s'estableix ser treballador per compte propi o d'altri com la categoria de referència i, per tant, es defineixen dues variables dicotòmiques: estar parat (SL1), que pren el valor 1 o 0 en cas contrari, i trobar-se en qualsevol altra situació excepte parat o treballant (SL2), que es codifica com a 1 o 0 si és un treballador per compte propi o d'altri o és a l'atur.

La component denominada **predictor lineal** d'aquest model és:

$$\beta_0+\beta_1 Edat_i+\beta_2 Dona_i+\beta_3 SL1_i+\beta_4 SL2_i+\beta_5 Ingressos_i$$

És important destacar el següent:

- En aquest exemple tenim sis paràmetres (els paràmetres beta), que són els que estimarem.
- Cada participant en aquest estudi de mercat té unes característiques pròpies d'edat, sexe, situació laboral i ingressos, i per això s'utilitza el subíndex *i*, que indica que s'utilitza el valor observat d'aquesta variable per a la *i*-èsima persona preguntada. Per això sol posar-se, al costat del model, la indicació *i* = 1-150, la qual cosa vol dir que hi ha un predictor lineal diferent per a cadascun dels participants en l'estudi. En general es parla d'*N* participants.
- Una vegada s'hagin estimat els paràmetres, es podrà calcular el predictor lineal per a qualsevol tipus de consumidor si se'n coneix l'edat, sexe, situació laboral i ingressos, encara que no n'hi hagi cap exactament com ell en l'estudi inicial.

La **component aleatòria** en aquest cas és una variable que pren dos valors possibles: «Sí compraria» o «No compraria». Se sol indicar amb la lletra *Y* per ser variable dependent, i el subíndex *i* per indicar que es refereix a l'*i*-èsim participant. Es pot codificar amb 1 en el cas afirmatiu i 0 en cas negatiu. S'estableix el següent comportament estocàstic (variable Bernoulli):

$$Y_i = \begin{cases} 1, & amb \quad probabilitat \quad p_i \\ 0, & amb \quad probabilitat \quad 1-p_i \end{cases}$$

Respecte a aquesta component, cal destacar que la probabilitat de comprar o no és individual de cada persona, i per aquest motiu s'utilitza una probabilitat amb el subíndex *i*. A més, per tractar-se d'una variable dicotòmica, el valor esperat de la variable correspon exactament a la probabilitat de comprar p_i. És a dir, l'esperança matemàtica $E(Y_i)$ es calcula així:

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

En definitiva, el valor que es predirà és la probabilitat que la variable dependent prengui el valor 1, és a dir $E(Y_i) = p_i = Prob(Y_i = 1)$.

La tercera component és una **funció de lligadura** que fa correspondre cada valor de p_i a un valor del predictor lineal i que compleix les condicions exigibles en el model lineal generalitzat. És a dir, ha de ser una funció monòtona, contínua i diferenciable.

L'elecció de la funció de lligadura determinarà finalment el tipus de model lineal generalitzat que s'hagi establert.

$$E(Y_i) = p_i = f(\beta_0 + \beta_1 Edat_i + \beta_2 Dona_i + \beta_3 SL1_i + \beta_4 SL2_i + \beta_5 Salari_i)$$

En el cas d'un model de regressió logística l'elecció seria la següent:

$$Prob(Y_i = 1) = \frac{\exp(\beta_0 + \beta_1 Edat_i + \beta_2 Dona_i + \beta_3 SL1_i + \beta_4 SL2_i + \beta_5 Ingressos_i)}{1 + \exp(\beta_0 + \beta_1 Edat_i + \beta_2 Dona_i + \beta_3 SL1_i + \beta_4 SL2_i + \beta_5 Ingressos_i)}.$$

Tots els models lineals generalitzats obeeixen a les tres parts considerades i, per tant, es determinen fixant:

1) quines són les variables que participen en la determinació del predictor lineal,

2) quina distribució estadística s'utilitza per a la variable dependent,

3) quina és la funció de lligadura.

2. Estimació per màxima versemblança

Per a procedir a l'estimació dels paràmetres, és necessari utilitzar un mètode que permeti optimitzar un criteri i que condueixi als millors valors possibles per a les «betes». La funció de versemblança es defineix a partir de les dades observades, del predictor lineal i la funció de lligadura, i del comportament estocàstic de la variable dependent. La funció de versemblança es basa en la independència de les observacions, i per tant significa que les respostes observades no tenen factors que les afectin i que puguin induir dependència entre elles. Per exemple, en el cas de l'anterior exemple de la perfumeria, no seria adequat que dues persones opinessin alhora sobre la fragància perquè la resposta d'una podria influir sobre la de l'altra.

Intuïtivament, la funció de versemblança és la que associa a cada vector de paràmetres possibles la probabilitat d'observar les dades disponibles si aquests paràmetres fossin certs. Per tant, en l'exemple que hem considerat, es tindria una funció de dimensió sis, el domini de la qual seria el producte de sis nombres reals i el seu recorregut l'interval [0,1], ja que el resultat és una probabilitat.

El procediment d'estimació d'un model lineal generalitzat consisteix a trobar els paràmetres que fan major la probabilitat final. El resultat és, doncs, un únic valor estimat per a cada paràmetre al qual s'associa un error estàndard.

La idea que hi ha al darrere de la maximització de la versemblança pot explicar-se metafòricament com la d'un explorador que busca les coordenades del cim d'una muntanya. Al cim hi ha el màxim i ha de trobar el punt de les coordenades del mapa on és aquest. Per a això inicia un recorregut a partir d'un lloc inicial i va ascendint per la muntanya fins a trobar un lloc en el qual ja no és possible pujar més amunt. En aquest moment, les coordenades donen el punt exacte del màxim.

El procediment iteratiu que van proposar Nelder i Wedderburn es basa en aquest principi i precisament la manera de definir els models lineals generalitzats garanteix que es pot buscar aquesta màxima amb poques iteracions.

3. Model de regressió logística o model lògit

El model de regressió logística especifica que:

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

On $X1$, ..., Xk es refereixen a les variables explicatives introduïdes en el model.

Però la inversa d'aquesta relació, que normalment s'anomena funció de lligadura (*link function*), és:

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

Anomenem $\frac{p_i}{(1 - p_i)}$ *odds* o quocient de probabilitats i, el seu logaritme, *log-odds*.

Per tant, en el model de regressió logística el predictor lineal és igual al logaritme del quocient de probabilitats. Com que el que es pretén predir és la probabilitat d'ocurrència d'algun succés, el model lògit és un model que permet predir la probabilitat d'aquesta ocurrència.

El model de regressió logística s'utilitza en molts àmbits: per exemple, en l'estudi de nous fàrmacs per a trobar la probabilitat que aquests medicaments siguin efectius en pacients amb determinades característiques; en economia per a predir si una acció incrementarà o reduirà el seu valor; en psicologia per a conèixer si la resposta a un estímul serà positiva o negativa; etc. El millor del model lògit és que la predicció sempre és entre 0 i 1 i per tant sempre és interpretable com una probabilitat o com un valor en percentatge. La funció lògit té una forma de *S*, de manera que si el predictor creix s'apropa a 1 i si decreix s'apropa a 0. Per tant, com més gran és el predictor més alta és la probabilitat, i quan el predictor és molt negatiu la probabilitat s'apropa a 0.

4. Interpretació de coeficients i *odds*-ràtios

Els paràmetres s'interpreten en el model de regressió logística binària d'una manera més complicada, en comparació de com s'interpreten en el context del model de regressió lineal.

Es necessita una mica de pràctica per a acostumar-se a parlar de probabilitats ajustades en lloc de valors ajustats. En interpretar els resultats de l'estimació d'un model de regressió logística, un primer pas útil és analitzar el signe dels paràmetres i comprovar si els signes estimats són els que la intuïció prèvia o la teoria indicaven.

Un paràmetre positiu significa que un augment en la variable que està associada a aquest paràmetre implica un augment en la probabilitat de la resposta del succés analitzat. Per contra, si un paràmetre és negatiu, quan el factor predictiu (variable independent) augmenta, la probabilitat del succés modelat disminueix. Per exemple, en el cas de l'exemple descrit anteriorment, si el paràmetre associat al salari fos positiu, s'interpretaria que a major salari major probabilitat que la persona es decideixi per comprar la fragància oferta pel fabricant. D'altra banda, si el paràmetre associat a l'edat fos negatiu, significaria que com més gran és la persona menys probable és que tingui intenció de comprar el producte.

Els models de regressió logística utilitzats en la pràctica sempre han de contenir un terme constant. El valor de la constant no és directament interpretable. S'utilitza com un valor mitjà i correspon al logaritme natural de la probabilitat del succés quan tots els regressors són igual a zero.

El quocient dels *odds* (anomenat *odds*-ràtio) és el nom donat a l'exponencial d'un paràmetre. Suposem que l'individu *i* canvia el seu *k*-èsim predictor, *Xk*, en una unitat. Per exemple, suposem que inicialment aquesta característica *Xk* és igual a *c* i llavors canvia per ser igual a $c + 1$. Llavors es veu fàcilment que:

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

Per tant,

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

Si la *j*-èsima variable explicativa és contínua, també podem deduir que el paràmetre β_j és el canvi proporcional en l'*odds*-ràtio o en l'elasticitat econòmica.

Un concepte connectat a l'*odds*-ràtio és la noció de raó de risc. La raó de risc és la proporció de probabilitats de successos quan l'indicador de predicció canvia en una unitat. Això és especialment important per a veure el canvi relatiu en la probabilitat d'un «succés» quan un indicador de risc és present en lloc d'absent. La definició és:

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

Les relacions de risc depenen del vector de variables explicatives i no poden expressar-se com una funció d'un sol paràmetre.

Suposem que en l'exemple de compra del perfum obtenim una estimació per al paràmetre de l'edat igual a -0,03. En ser negatiu ja podem afirmar que com més edat menys probabilitat de comprar el producte, essent la resta de característiques les mateixes. L'*odds*-ràtio seria igual a *exp*(-0,03) = 0,97. Això vol dir que quan augmenta un any l'edat de la persona la probabilitat de comprar es multiplica per 0,97, la qual cosa vol dir que disminueix el 3%.

5. Matriu de confusió

Les estimacions de paràmetres s'indiquen amb un accent ^ damunt de la beta. Una vegada obtinguts, podem predir la probabilitat del succés modelat per la resposta. Donades unes característiques, podem calcular la probabilitat ajustada, que es basa en l'expressió següent:

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

Per a construir la **matriu de confusió** utilitzem els casos del conjunt de dades original (que pot denominar-se conjunt d'aprenentatge o *training set*). Per a cada cas, es té una resposta observada i es pot predir la probabilitat ajustada pel model. La matriu de confusió permet comparar aquests dos valors. Per exemple, un individu en el conjunt de dades pot tenir una resposta igual a 1, i la probabilitat estimada pel model pot ser 0,9 o, el que és el mateix, 90%. Aquest cas és bastant plausible. Però pot ocórrer que la resposta sigui 1, mentre que la predicció del model sigui molt baixa, diguem del 20%. En aquest cas, la predicció del model seria sorprenent.

En el nostre exemple, considerem una persona que tria comprar la fragància però que té unes característiques totalment allunyades dels consumidors més propensos. Això li donaria una probabilitat estimada de compra de solament 20%. Podem dir que la probabilitat que compri el producte és d'1 contra 4. En la majoria dels conjunts de dades, quan es comparen prediccions i observacions, podem trobar casos concrets en els quals l'observació i la probabilitat previstes no són concordants. Aquests són errors del model. Pot ser que es doni el cas contrari al que s'acaba de comentar, una molt alta probabilitat de triar la resposta afirmativa; no obstant això, s'observa el contrari.

En models de regressió lineal, era raonable discutir la correlació entre observacions i prediccions en termes d'estadístic R^2, el coeficient de determinació. No obstant això, en models generalitzats, les correlacions manquen del mateix sentit.

En un model de regressió logística, quan s'observa la probabilitat predita pel model, no se sap si el valor és massa alt o massa baix. Normalment, un bon llindar de discriminació és el de 50%. Això correspon a un *odds* igual a 1, la qual cosa significa que les dues opcions de resposta són equiprobables. Fixat aquest llindar, es considera que si la resposta és positiva i la probabilitat estimada és superior a 50%, llavors el model ha encertat. D'igual manera, si la resposta és negativa i la probabilitat és inferior a 50%, el model també ha aconseguit

encertar. No obstant això, quan la probabilitat no es correspon amb l'observat es tracta de casos en els quals el model no encerta. Vegem un exemple de taula de confusió:

Taula 3. Exemple de matriu de confusió. Nombre de casos

	Probabilitat predita inferior a 50%	**Probabilitat predita superior o igual a 50%**	**Total**
Observat: Sí	50	10	60
Observat: No	30	60	90
Total	80	70	150

Del total de 150 casos, veiem que 110 casos estan correctament classificats, ja que el model prediu la resposta correcta per als 50 participants que sí comprarien el producte i tenen una probabilitat predita elevada, i per als 60 que no el comprarien i tenen una probabilitat baixa. Això significa que el percentatge general de classificació correcta és igual a 110/150 = 73,33%, la qual cosa és excel·lent. No obstant això, hi ha un nombre de casos que no tenen la resposta esperada. Per exemple, hi ha 30 casos que no trien «comprar el producte», però la probabilitat és alta i el model prediu que és probable que hagin fet aquesta elecció. D'altra banda, hi ha 10 casos que responen que comprarien el producte però el model prediu que la seva probabilitat de compra és inferior a 50%.

Perquè un model de regressió logística amb finalitats predictives tingui èxit, el nombre de casos que es classifiquen correctament ha de ser alt, mentre que el nombre de casos que es classifiquen incorrectament ha de ser baix.

El nombre de resultats denominats falsos positius correspon a casos en què la predicció de la probabilitat de la resposta afirmativa és elevada però la resposta observada és negativa. En aquest exemple hi ha 30 falsos positius. Els falsos positius no són bons en la pràctica. En el nostre exemple, signifiquen que el model prediu que el client és probable que vagi a comprar, per la qual cosa és possible que s'hagi fet una campanya publicitària per arribar a aquest col·lectiu; no obstant això, no es produeix la resposta desitjada. Els falsos positius poden conduir, en aquest exemple, al fracàs dels esforços comercials. Si usem el model per a predir el comportament d'aquests clients i tractem de vendre'ls la «nova fragància» sense èxit, incorrerem en uns costos publicitaris que podrien ser estalviats.

El nombre de respostes corresponents als falsos negatius correspon al nombre de casos on el model prediu que el client té una probabilitat de compra baixa, i no obstant això els participants sí han triat comprar. En el nostre exemple, els clients que han triat comprar la fragància mentre la predicció de la probabilitat de fer-ho és baixa, inferior a 50%, corresponen a un total de 10 casos.

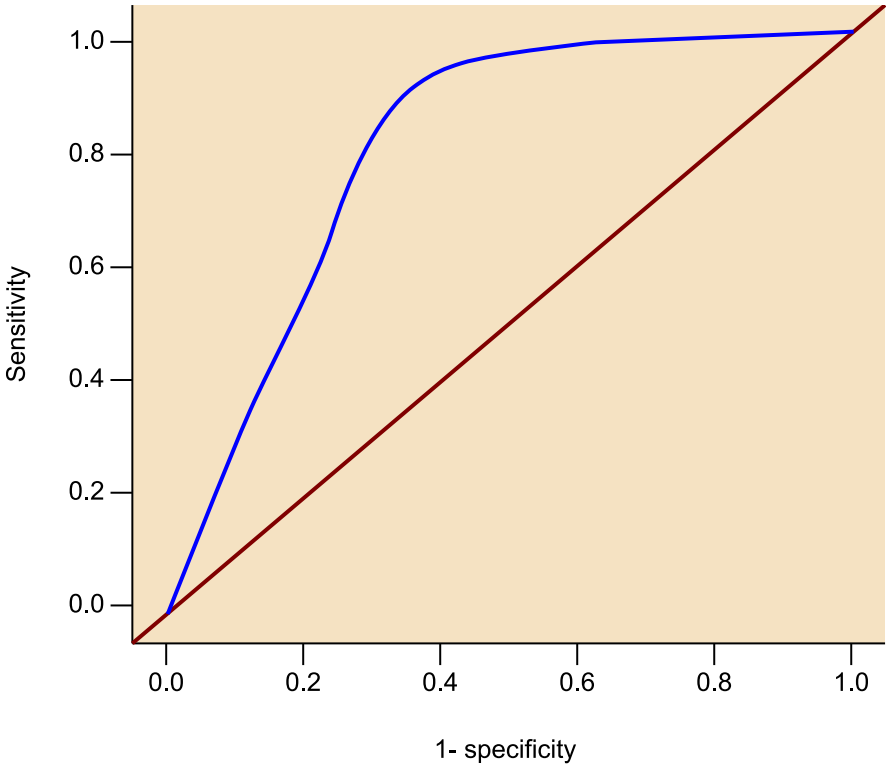
Aquest model té solament uns pocs falsos negatius. I si s'utilitzés amb finalitats predictives, significa que solament hi hauria uns quants clients que, àdhuc no responent al perfil del comprador, comprarien el producte.

Idealment, un model perfecte en termes de capacitat predictiva no tindria ni falsos positius ni falsos negatius en la taula de confusió. Però hi ha altres mesures que també solen tenir-se en compte. La sensibilitat és la proporció dels classificats correctament entre els veritables participants que han donat resposta afirmativa. L'especificitat és la proporció de casos correctament classificats entre les respostes negatives. En la taula 3, la sensibilitat és 50 / (50 + 10) = 83,33% i l'especificitat és 60 / (60 + 30) = 66,67%.

6. Corba ROC. Mètriques AUROC i KS

Hi ha una eina gràfica per a avaluar la bondat d'ajust en la regressió logística. Presentem la corba *Receiver Operating Characteristic*, també coneguda com la corba ROC. La corba ROC és un gràfic de la sensibilitat davant d'1 menys l'especificitat. Cada punt en la corba correspon a un nivell llindar de discriminació en la matriu de confusió. És a dir, així com en l'exemple anterior s'havia considerat un llindar de 50%, es construeixen totes les matrius canviant aquest llindar des d'1% fins a 99%, i es va calculant la sensibilitat i 1 menys l'especificitat. El millor model en termes d'ajust seria el que tingués una corba ROC al més a prop possible de la cantonada superior esquerra de la gràfica. Un model no discriminant tindria una corba ROC plana, prop de la diagonal. L'anàlisi ROC proporciona una manera de seleccionar models possiblement òptims i subòptims basada en la qualitat de la classificació a diferents nivells o llindars. Per tenir una regla objectiva de comparació de les corbes ROC, es calcula l'àrea situada entre la corba i la diagonal, anomenada simplement AUROC (*Area Under the ROC*). El model l'àrea del qual sigui superior és el preferit. Una altra manera de mesurar la qualitat de l'ajust del model és mitjançant la prova de Kolmogorov-Smirnov, en la qual es compara la distribució de les probabilitats estimades amb el model en els dos grups de resposta considerats. Si és diferent, es considera que el model predictiu discrimina bé.

Figura 1: Exemple de corba ROC



7. Selecció de llindar i altres models

Un dels elements pràctics més importants és saber seleccionar el llindar de classificació (*threshold*) més adequat perquè el model sigui útil des d'un punt de vista predictiu. Això vol dir que, potser, no és el més adequat fixar un punt de tall en 50%. Per a això, convé fixar-se en aquell punt que proporciona una major distància entre la corba ROC i la diagonal. Aquest nivell de llindar de classificació és el que dona una major sensibilitat i especificitat en el model, és a dir, una major capacitat de classificar correctament els participants en l'estudi.

A continuació s'expliquen els dos models, model pròbit i model de Poisson:

1) Model pròbit. El model pròbit s'utilitza per a casos en els quals la resposta és dicotòmica i es vol una alternativa al model de regressió logística. Encara que hi ha una equivalència entre els paràmetres de tots dos models demostrada per Amemiya (1981), si es divideixen per 1,6 els coeficients del model lògit, s'obtenen els del model pròbit; encara hi ha vegades en què es prefereix el model lògit perquè és millor per a casos en els quals hi ha més observacions extremes i es vol insistir en la interpretació dels *odds*.

En el model pròbit, l'especificació és la següent:

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

2) Model de Poisson. El model de Poisson s'utilitza per a casos en els quals la variable dependent és el nombre de vegades en les quals ocorre algun succés. S'especifica així:

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

En aquest cas se suposa que la distribució de probabilitat de Y_i és una distribució de Poisson. La interpretació dels paràmetres es fa en termes del seu signe. Quan un paràmetre és positiu, això vol dir que si la seva característica associada augmenta, llavors també ho fa el nombre esperat de vegades que té lloc el fenomen analitzat. En cas contrari, si el paràmetre és negatiu, llavors en augmentar aquesta variable disminueix el nombre esperat que s'està modelitzant.

Model de Poisson

Un exemple concret de model de Poisson és el de veure quantes nits pernocten els turistes que arriben a una ciutat. Si s'obté que l'edat té un coeficient positiu, llavors s'espera que com més edat més elevat sigui el nombre de nits que s'espera que passin els turistes a la ciutat. Si, per exemple, el coeficient associat a una nacionalitat fos negatiu, s'interpretaria

que els turistes d'aquesta nacionalitat s'espera que passin menys nits a la ciutat que els que provenen d'altres països, suposant que tinguessin la resta de característiques iguals.

Bibliografia

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

Guillén, M. (2014). «Regression with Categorical Dependent Variables». A: Frees, E. W.; Derrig, R.; Meyer, G. (ed.). *Predictive Modeling Applications in Actuarial Science*. Vol. 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), pàg. 255-265).

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

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

Nelder, J. A.; Baker, R. J. (1972). «Generalized linear models». *Encyclopedia of statistical sciences*.

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

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

Enllaç d'internet

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