Universidad de Granada

Facultad de Ciencias

Regresión Logística

Análisis Exploratorio de Datos

Patricia Contreras Parra

Departamento de Estadística e Investigación Operativa
Granada, 2023

Índice

1. Introducción

Los problemas de clasificación son frecuentes, incluso más que los de regresión. Algunos ejemplos son,

1. Una persona llega al servicio de urgencias con una serie de síntomas que podrían atribuirse a una de tres infecciones. ¿Cuál de las tres padece el individuo?

2. Un servicio de banca en línea debe ser capaz de determinar si una transacción que se está realizando en el sitio es fraudulenta basándose el historial de transacciones anteriores.

3. Un estudio quiere establecer un modelo que permita calcular la probabilidad de obtener una matrícula de honor al final del bachillerato en función de la nota que se ha obtenido en matemáticas. La variable matrícula está codificada como 0 si no se tiene matrícula y 1 si se tiene.

Al igual que en la regresión, en la clasificación se tiene un conjunto de observaciones de entrenamiento $(x_1, y_1), ..., (x_n, y_n)$ que podemos utilizar para construir un clasificador. Queremos que nuestro clasificador funcione bien no sólo con los datos de de entrenamiento, sino también en las observaciones de prueba que no se utilizaron para entrenar el clasificador.

2. Antecedentes

2.1. Clasificación

El modelo de regresión lineal supone que la variable de respuesta Y es cuantitativa. Pero en muchas situaciones, la variable respuesta es *cualitativa*. Por ejemplo, el color de los ojos es cualitativa. En ocasiones, las variables cualitativas se denominan categóricas, estos términos los utilizaremos indistintamente. El proceso por el que se predice respuestas cualitativas es conocido como *clasificación*. Predecir la clasificación de una respuesta cualitativa para una observación puede denominarse como el proceso de *clasificar* una observación, ya que implica asignar la observación a una categoría o clase. Por otra parte, los métodos de clasificación suelen predecir en primer lugar la probabilidad de que la observación pertenezca a cada una de las categorías de una variable cualitativa, como base para realizar la clasificación. En este sentido, también se comportan como métodos de regresión.

Hay muchas técnicas de clasificación posibles o *clasificadores*, que se pueden utilizar para predecir una respuesta cualitativa, como la *regresión logística, el análisis discriminante lineal, el análisis discriminante cuadrático, Naive Bayes* y *los vecinos más cercanos a K.*

2.2. ¿Por qué no a la Regresión Lineal?

La regresión lineal, en general, no es apropiada en el caso de predecir una respuesta cualitativa. ¿a qué se debe esta afirmación?

El modelo de regresión lineal simple es utilizado para predecir una respuesta cuantitativa Y a partir de una única variable predictora X y de ellas se supone que existe una relación lineal. Dicho modelo adopta la siguiente forma:

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{2.1}$$

Este modelo se construye a partir de una combinación lineal de las variables. En este caso, la variable Y es numérica, continua y no acotada. Ejemplo puede ser, la temperatura, los ingresos, etc. En el caso en el que la variable dependiente Y es cualitativa binaria (dos niveles) esta adoptaría la siguiente forma:

$$Y = \begin{cases} 0 & \text{No evento} \\ 1 & \text{Evento} \end{cases}$$

En la que Y toma el valor 1 si ocurre un evento o 0 en el caso contrario. Ejemplos clásicos de ello son: si se aprueba o no un préstamo bancario, se tiene cáncer o no, etc. Para una respuesta binaria como la anterior, la regresión por mínimos cuadrados $Y = \beta_0 + \beta_1 X$ no es del todo descabellada. Sin embargo, si utilizamos

la regresión lineal, algunas de nuestras estimaciones podrían estar fuera del intervalo [0, 1] (véase la Figura 2.2), lo que las haría difícil de interpretar como probabilidades. El objetivo ahora es, tratar con un modelo que prediga la variable cualitativa binaria Y sabiendo que esta predicción se basa en la ocurrencia de un evento.

A continuación, mediante dos ejemplos prácticos mostramos como un modelo de regresión lineal simple no siempre se ajusta bien a los datos según qué tipo de variable se pretende predecir. El primer conjunto de datos pretende explicar la resistencia de una soldadura en función de su edad. El segundo conjunto de datos procede de `Default` en `R`, nos interesa predecir si un individuo incumplirá el pago de su tarjeta de crédito (`default`) en función del saldo medio que le queda en su tarjeta de crédito después de efectuar el pago mensual (`balance`).

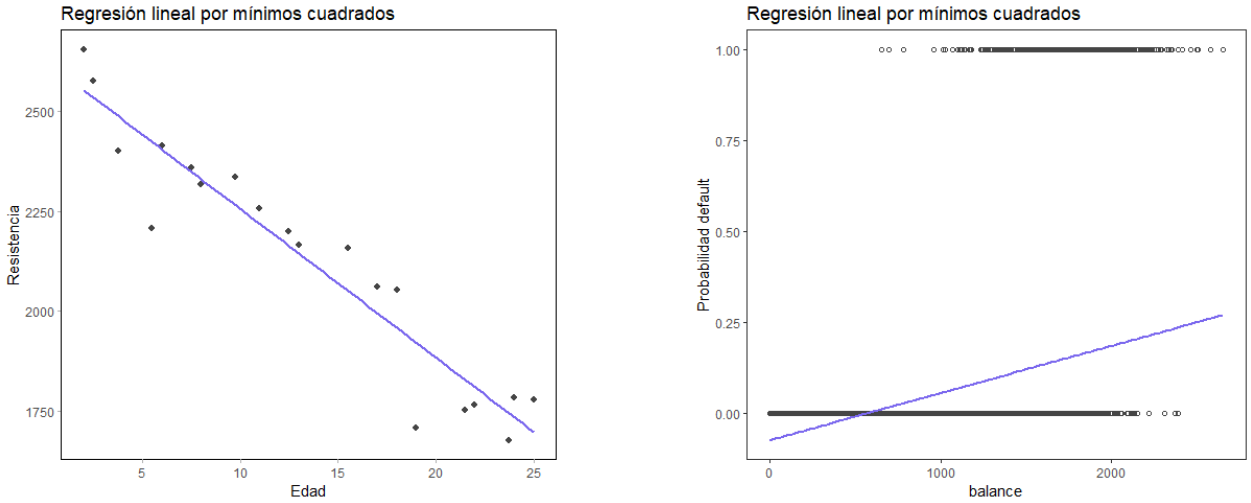

FIGURA 2.2 *Ejemplo Regresión Lineal.* Izquierda: *Predicción de la resistencia de una soldadura en función su edad usando regresión lineal. La línea se ajusta a la escala de los datos.* Derecha: *Predicción estimada de probabilidad de* `default` *usando regresión lineal. Algunas probabilidades son negativas. Las marcas grises indican los valores 0/1 codificados de default (No o Sí).*

Es preferible utilizar un método de clasificación que sea realmente adecuado para los valores de respuesta cualitativos. En la siguiente sección, presentamos la regresión logística, que se adapta bien al caso de una respuesta cualitativa binaria.

3. Regresión Logística

En la sección 2.2 considerábamos la posibilidad de utilizar un modelo de regresión lineal para representar esta probabilidad:

$$Pr(Y = 1|X) = p(X) = \beta_0 + \beta_1 X \tag{3.1}$$

Esta sería la probabilidad condicionada de que un dato pertenezca a la categoría cero o a la categoría uno, dadas unas variables de entrada X. Si volvemos al ejemplo de los datos `Default`, predecir por un modelo lineal la Pr(`default=sí`) utilizando p(`balance`), obtenemos el modelo mostrado en el panel derecho de la Figura 2.2. El problema de este enfoque es el siguiente: para saldos cercanos a cero predecimos una probabilidad de impago negativa; si tuviéramos que predecir para saldos muy grandes, obtendríamos valores superiores a *1*. Estas predicciones no son sensatas, ya que por supuesto, la verdadera probabilidad de impago, independientemente del saldo de la tarjeta de crédito, debe situarse entre *0* y *1*. El problema no es exclusivo de los datos de impago de créditos. Cada vez que una línea recta corresponde a una respuesta binaria codificada como *0* ó *1*, en principio siempre podemos predecir $p(X) < 0$ para algunos valores de X y $p(X) > 1$.

La idea de usar la probabilidad de que ocurra un evento como variable de respuesta a través de un modelo lineal no parece ser la más adecuada ya que se requiere "forzar" a que los valores predichos de la probabilidad se encuentren entre *0* y *1*, independientemente de los valores que tome la variable de entrada X. Además de que la mejor relación entre ambas variables sería ser representadas por una curva sigmoidal. Para lograr ambas condiciones, dejando de lado el modelo lineal, se consigue mediante una transformación logística de la $p(X)$ de que ocurra un evento, consistente en aplicar el logaritmo al *odds* asociado a esta probabilidad. El modelo sería $Y = \beta_0 + \beta_1 X$, pero definiendo Y del modo siguiente,

$$Y = ln\left(\frac{p(X)}{1 - p(X)}\right) \tag{3.2}$$

Lo que es la denominada la función *logit* o *log odds*. Los *odds* de un suceso, es el cociente de la probabilidad de ocurrencia de un evento entre su probabilidad de no ocurrencia y se nota como: $[p(X)/[1 - p(X)]$. Puede tomar cualquier valor comprendido entre 0 y $+\infty$. Los valores de las probabilidades cercanos a 0 y $+\infty$ indican probabilidades muy bajas y muy altas, respectivamente. Por ejemplo, si un paciente tuviese la probabilidad de desarrollar una enfermedad coronaria igual a $p(X) = 0{,}679$, entonces el *odds* sería, $\frac{0{,}679}{1-0{,}679} = 2{,}12$, es decir, desarrollar una enfermedad coronaria es 2.12 veces más probable que no desarrollarla. Tradicionalmente, en las carreras de caballos se utilizan los *odds* en lugar de las probabilidades, ya que se relacionan de forma más natural con la estrategia de apuesta correcta.

Siguiendo con la expresión (3.2), esta es igual a,

$$ln\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X \tag{3.3}$$

Este sería nuestro *modelo de regresión logístico binario*. El término a la derecha de la igualdad es la expresión de una recta, idéntica a la del modelo general de regresión lineal. Este modelo está conceptuado como el recurso más eficiente para representar el vínculo entre una variable binaria de respuesta y la variable independiente. El modelo (3.3) es equivalente a,

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X} \tag{3.4}$$

De forma que si queremos conocer el valor de $p(X)$, nos queda despejar y las siguientes expresiones son equivalentes,

$$Pr(Y = 1|X) = p(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \tag{3.5}$$

Esta última expresión (3.5) es la conocida *función logística* y es lo que usa la regresión logística para modelar la probabilidad de que ocurra un evento. Más tarde la utilizaremos para estimar los coeficientes del modelo. En una representación de la curva *logit* (3.2), el eje X tiene como valores posibles el rango *0* y *1* como se puede observar en el panel izquierdo de la Figura 3. Como queremos que dichos valores se encuentren en el eje Y, es necesario obtener la inversa de la función *logit*. La inversa de la función *logit* es la *función logística* (3.5).

El panel derecho de la Figura 3 ilustra la probabilidad del modelo de regresión logística a los datos `Default`. Obsérvese que para saldos bajos (`balance`), ahora predecimos la probabilidad de impago como cercana, pero nunca por debajo de cero. Del mismo modo, para saldos elevados, predecimos una probabilidad de impago cercana, pero nunca superior, a uno. La función logística siempre producirá una curva en forma de "S", por lo que, independientemente del valor de X, obtendremos una predicción sensata.

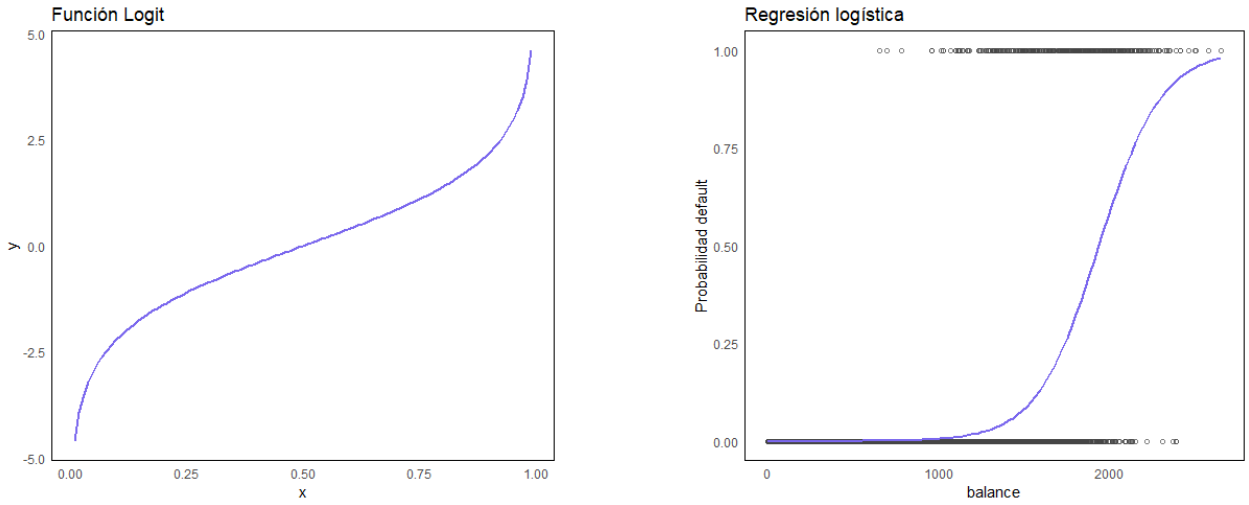

FIGURA 3 Izquierda: *Función Logit: Simulación de datos entre (0,1).* Derecha: *Modelo de Regresión Logística ajustado a los datos de* `Default`.

3.1. Interpretación de los Coeficientes

El componente central de la regresión logística es el *odds* de que se produzca un evento, es decir,

$$\frac{p(X)}{1-p(X)}$$

Puesto que, $ln\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$, la regresión logística modela linealmente el logaritmo del *odds*. Si consideramos dos sujetos con valores x_1 y x_2 de la variable X, según el modelo considerado, para el primer sujeto tendremos: $ln\left(\frac{p(x_1)}{1-p(x_1)}\right) = \beta_0 + \beta_1 x_1$ y, para el segundo, $ln\left(\frac{p(x_2)}{1-p(x_2)}\right) = \beta_0 + \beta_1 x_2$ donde $p(x_1)$ es la probabilidad de que un individuo con valor x_1 de la variable X presente la característica de interés y $p(x_2)$ representa la misma probabilidad con valor x_2 de la variable X. Si restamos estas dos igualdades tenemos la diferencia de ambas,

$$ln\left(\frac{p(x_1)}{1-p(x_1)}\right) - ln\left(\frac{p(x_2)}{1-p(x_2)}\right) = (\beta_0 + \beta_1 x_1) - (\beta_0 + \beta_1 x_2) = \beta_1 x_1 - \beta_1 x_2 \tag{3.6}$$

Aplicando la propiedad de los logaritmos llegamos a,

$$ln\left[\frac{\frac{p(x_1)}{1-p(x_1)}}{\frac{p(x_2)}{1-p(x_2)}}\right] = \beta_1(x_1 - x_2) \tag{3.7}$$

El cociente al que se le aplica el logaritmo es una razón entre dos *odds*, de modo que no es otra cosa que un *odds ratio* o la razón de los *odds*, que vamos a representar por el símbolo *OR*. La expresión anterior (3.7) se puede expresar como,

$$ln(OR) = \beta_1(x_1 - x_2) \tag{3.8}$$

lo que es igual a,

$$OR = e^{\beta_1(x_1 - x_2)} \tag{3.9}$$

Todo esto se puede resumir diciendo que la razón de *odds* entre dos individuos con valores x_1 y x_2 de la variable independiente se puede conseguir elevando el número e al producto $\beta_1(x_1 - x_2)$. Si consideramos el caso particular en el que $x_1 = x_2 + 1$ (si los valores de X se diferencian en una unidad), tendremos:

$$ln(OR) = \beta_1(x_1 - (x_1 - 1)) = \beta_1 \tag{3.10}$$

De modo que β_1 se puede interpretar como el logaritmo de la razón de *odds* de presentar la característica para dos individuos que se diferencia en una unidad respecto a la variable independiente. Aplicando la función exponencial, esta última expresión adopta la siguiente forma,

$$OR = e^{\beta_1} \tag{3.11}$$

Por lo que, el exponencial de β_1 no es más que el *odds ratio* entre dos individuos que se diferencian en una unidad de la variable independiente. Entonces, en un modelo de regresión logística, la cantidad que cambia $Pr(Y = 1|X)$ debido a un cambio de una unidad en X depende del valor actual de X. Pero independientemente del valor de X, si,

- $\beta_1 > 0$, la curva logística es creciente, al aumentar la X aumenta la probabilidad de $Y = 1$.
- $\beta_1 = 0$ la probabilidad de $Y = 1$ no depende de X.
- $\beta_1 < 0$, la curva logística es decreciente, al aumentar el valor de la X disminuye la probabilidad de que $Y = 1$.

3.2. Estimación de los Coeficientes de Regresión Logística

Los coeficientes β_0 y β_1 son desconocidos y debemos estimarlos. Aunque podríamos utilizar mínimos cuadrados (no lineales) para estimar el modelo, el método más adecuado es por *máxima verosimilitud* ya que tiene mejores propiedades estadísticas.

El procedimiento para la *estimación máximo verosímil* sería considerar una muestra de tamaño N, con etiquetas *0* ó *1*. Matemáticamente, para las muestras etiquetadas como *1*, intentamos estimar β de modo que el producto de todas las probabilidades $p(x)$ sea lo más cercano posible a *1*. Y para las muestras etiquetadas como *0*, intentamos estimar β de manera que el producto de todas las probabilidades sea lo más cercano posible a *0*; en otras palabras, $(1 - p(X))$ debe ser lo más cercano posible a *1*. La intuición anterior se representa como:

- Para los elementos etiquetados como *1*: $\prod_{i:y_i=1}^{N} p(x_i)$
- Para los elementos etiquetados como *0*: $\prod_{i':y'_i=0}^{N}(1 - p(x_{i'}))$

Al combinar las condiciones anteriores, queremos encontrar los parámetros β_0 y β_1 tales que el producto de ambos productos sea máximo en todos los elementos del conjunto de datos.

$$L(\beta_0, \beta_1) = \prod_{i:y_i=1}^{N} p(x_i) \prod_{i':y'_i=0}^{N} (1 - p(x_{i'})) \tag{3.12}$$

Combinamos los productos y tomamos la probabilidad logarítmica para simplificarlo aún más,

$$L(\beta_0, \beta_1) = \prod_{i}^{N} (p(x_i)^{y_i} \cdot (1 - p(x_i))^{1-y_i}) \tag{3.13}$$

$$\ell n(\beta_0, \beta_1) = \sum_{i=1}^{N} y_i ln(p(x_i)) + (1 - y_i) ln(1 - p(x_i))$$

Sustituimos $p(X)$ por su forma exponente,

$$\ell n(\beta_0, \beta_1) = \sum_{i=1}^{N} y_i ln\left(\frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}}\right) + (1-y_i) ln\left(1 - \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}}\right) \tag{3.14}$$

$$\ell n(\beta_0, \beta_1) = \sum_{i=1}^{N} y_i \left[ln\left(\frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}}\right) - ln\left(\frac{e^{-(\beta_0+\beta_1 x_i)}}{1+e^{-(\beta_0+\beta_1 x_i)}} -\right)\right] + ln\left(\frac{e^{-(\beta_0+\beta_1 x_i)}}{1+e^{-(\beta_0+\beta_1 x_i)}}\right)$$

Aplicamos las propiedades de los logaritmos y simplificando nos queda,

$$\ell n(\beta_0, \beta_1) = \sum_{i=1}^{N} y_i \left[ln\left(e^{\beta_0+\beta_1 x_i}\right)\right] + ln\left(\frac{1}{1+e^{(\beta_0+\beta_1 x_i)}}\right) \tag{3.15}$$

Finalmente, la función que se optimizaría, sería,

$$\ell n(\beta_0, \beta_1) = \sum_{i=1}^{N} y_i(\beta_0 + \beta_1 x_i) - ln\left(1 + e^{(\beta_0+\beta_1 x_i)}\right) \tag{3.16}$$

y se maximiza la función. Existen muchos métodos para hacerlo como el Método de Newton-Raphson o el método de Müler.

	Coefficients	Std.error	z-statistic	p-value
`Intercept`	10.6513	0.3612	29.5	< 0.0001
`balance`	0.0055	0.0002	24.9	< 0.0001

TABLA 3.2 *Datos de* `Default`: *coeficientes estimados del modelo de regresión logística que predicen la probabilidad de impago* (`default`) *utilizando el saldo* (`balance`). *Un aumento de una unidad del saldo* (`balance`) *se asocia a un aumento de log odds de 0,0055 unidades.*

La tabla 3.2 muestra las estimaciones de coeficientes que resultan de aplicar un modelo de regresión logística a los datos `Default` para predecir la probabilidad de Pr(`default=sí`) utilizando el saldo (`balance`). Vemos que $\hat{\beta}_1 = 0,0055$, lo que indica que un incremento del saldo (`balance`) está asociado a un incremento de la probabilidad de impago (`default`). Para ser precisos, un aumento de una unidad en el saldo (`balance`) se asocia con un aumento de la probabilidad logarítmica de impago (`default`) de 0,0055. unidades.

3.3. Convertir Probabilidad en Clasificación

Para conseguir la clasificación en un modelo de regresión logística, es necesario establecer un umbral (*threshold*) de probabilidad a partir del cual, se considera que la variable pertenece a uno de los niveles, *0* y *1*. Por ejemplo, se puede asignar una observación al grupo *1* si $\hat{P}(Y = 1|X) > 0,5$ y al grupo *0* si de lo contrario.

En el caso de los datos de `Default` se puede decir que el banco clasificará como clientes de alto riesgo a aquellos que tengan una probabilidad de entrar en incumplimiento (`default`) mayor a 0,5 (podría ser más exigente y poner como límite 0,1).

3.4. Predicciones del Modelo de Regresión Logística

Una vez estimados los coeficientes del modelo logístico, es posible conocer la probabilidad de que la variable dependiente pertenezca al nivel de referencia, dado un determinado valor del predictor. Para ello se emplea la ecuación del modelo:

$$\hat{Pr}(Y = 1|X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} \tag{3.17}$$

Por ejemplo, utlizando las estimaciones del coeficiente que figuran en la tabla 3.2, predecimos que la probabilidad de impago para un cliente con un saldo de 1,000$ es de,

$$\hat{P}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10{,}6513+0{,}0055\times 1000}}{1+e^{-10{,}6513+0{,}0055\times 1000}} = 0{,}00576$$

Dado un cliente con un saldo de 1,000$, el modelo de regresión logística estima que la probabilidad de que el cliente entre en incumplimiento es de 0,576 %. Esto implica que, si el banco ha establecido un umbral de probabilidad en el que, aquellos que tengan una probabilidad mayor a 0,5 entran en el grupo de clientes que sí incumplen sus pagos, según el resultado obtenido, es bastante improbable que el cliente tenga un inclumpliento de pago y este cliente se clasifícaría en el grupo *0*.

4. Condiciones del Modelo

La regresión logística no requiere de ciertas condiciones como linealidad, normalidad y homocedasticidad de los residuos que sí lo son para la regresión lineal. Las principales condiciones que requiere este modelo son:

- **Respuesta binaria:** La variable dependiente ha de ser binaria.
- **Independencia:** las observaciones tienen que ser independientes unas de otras.
- **Relación lineal entre el logaritmo natural de odds y la variable continua.**
- **La regresión logística** no precisa de una distribución normal de la variable continua independiente.
- **Número de observaciones:** no existe una norma establecida al respecto, pero se recomienda entre 50 a 100 observaciones.

5. Construcción y Evaluación de Modelos

Siempre que se ajusta un modelo de regresión de cualquier tipo, una precaución importante a los efectos de sacar conclusiones es la de corroborar que este modelo representa adecuadamente el proceso que se estudia, y por ende, que sea compatible efectivamente con los datos usados. Diferentes áreas que deben ser consideradas en el proceso de conformación de modelos en regresión logística son descritas en este apartado. En la primera sección se tratan las definiciones de *lejanía* y de *razón de verosimilitud*, nociones que permiten evaluar características que son deseables que cumplan en modelo que se ajusta, seguidamente trataremos con el estadístico de *Wald* para el constraste de los parámetros del modelo, abordaremos la evaluación de bondad del ajuste y por último daremos una pequeña pincelada sobre los residuos del modelo.

5.1. Lejanía y Razón de Verosimilitud

En el contexto de regresión logística, la expresión de *verosimilitud* del modelo que vimos en el apartado 3.2. no es más que una medida razonable para valorar el grado en el que el modelo arroja resultados coherentes con la condición que tienen los sujetos de la muestra. Recordemos,

$$V = \prod_i^N (p(x_i)^{y_i} \cdot (1-p(x_i))^{1-y_i})$$

Utilizaremos entonces la verosimilitud para valorar la capacidad predictiva del modelo. La siguiente expresión,

$$L = -2ln(V) \tag{5.1}$$

Es conocida como *lejanía* del modelo o *deviance*. Dado que la verosimilitud es un producto de probabilidades, el valor resultante de $V < 1$, por el logaritmo siempre será negativo, de modo que la lejanía siempre será un número positivo.

Sin embargo, antes de aceptar un modelo, debemos tener en cuenta que aún existe un modelo más simple en la regresión logística y este es *el modelo nulo*, es decir, aquel que,

$$ln\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 \tag{5.2}$$

Es decir, la variable Y no depende de la variable independiente. En este caso, la estimación máximo verosímil de $p(X)$ sería el cociente entre casos favorables y caso posibles. Siempre que se compute el modelo, el algoritmo de la regresión logística ejecuta dos lejanías: la que corresponde propiamente al modelo que se ha ajustado (L), y la que corresponde al *modelo nulo*, (L_0).

La lejanía del modelo nulo es más grande que la de cualquier modelo ampliado. Esto es razonable debido a que se trata de un modelo mucho menos sofisticado y debe necesariamente tener una falta de ajuste mayor. La diferencia entre estas lejanías mide "el aporte" que hacen las variables incorporadas al modelo. Es decir, para valorar dicho aporte se puede calcular,

$$RV = L_0 - L = -2ln(V_0) + 2ln(V) = -2ln\left(\frac{V_0}{V}\right) \tag{5.3}$$

Este estadístico es conocido como la *razón verosimilitudes* y se distribuye según una *Chi- cuadrado* con *1* grados de libertad, χ^2_1. RV sirve para evaluar si la variable X tomada en el conjunto contribuye efectivamente a explicar la modificación que se producen en $p(X)$. En el caso de los datos de `Default`, la lejanía del modelo para predecir el impago en función del saldo, es,

$$-2ln(V) = 1596{,}5$$

y la del modelo nulo, igual a,

$$-2ln(V_0) = 2920{,}6$$

Aplicando el resultado (5.4), la razón de verosimilitudes sería,

$$RV = 1324{,}198$$

Dado que el percentil 95 de la distribución χ^2 con 1 grado de libertad es igual a 3.84, se rechaza la hipótesis de que el modelo nulo es sufiente y concluimos que la probabilidad de impago se explica mejor con la variable de saldo.

5.2. Contraste sobre los Parámetros del Modelo

El *estadístico de Wald* (z-statistic) proporciona la contribución individual de cada uno de los parametros del modelo. Para contrastar si el parametro β_1 es significativo o no, se tendrá que realizar el test de significación de parámetros,

$$\begin{cases} H_0: & \beta_1 = 0 \\ H_1: & \beta_1 \neq 0 \end{cases}$$

El contraste evalúa si el parámetro asociado a una variable es igual a cero o no. Para el parámetro β_1, el estadístico de Wald sería,

$$Z_{Wald} = \frac{\hat{\beta}_1}{S.E.(\hat{\beta}_1)} \tag{5.4}$$

donde $\hat{\beta}_1$ es la estimación máximo verosímil de β_1 y $S.E.(\hat{\beta}_1)$ es su error estándar. Se verifica que Z sigue una distribución estándar, o lo que es equivalente, $Z^2 \rightsquigarrow \chi^2_1$. Por lo tanto, se rechazará la hipótesis nula a un nivel de significación cuando se verifique que, el valor del estadístico es superior a el valor de $\chi^2_{1;\alpha}$. Así,

asumimos que dicha variable predictora está haciendo una contribución significativa al modelo para predecir el valor de Y y por tanto se debe mantener. En el caso de los datos de `Default`, el estadístico para $\hat{\beta}_1$ sería,

$$Z_{Wald} = \frac{0{,}0055}{0{,}0002} = 24{,}95$$

El percentil 95 de la distribución χ^2 con 1 grado de libertad es igual a 3.84. La conclusión es la misma. La variable saldo es significativa para explicar la variable Y.

La razón de verosimilitudes y la prueba de Wald brindan resultados similares en muestras grandes, por ello, en estudios con suficiente tamaño de muestra no importa cuál de las dos pruebas sea la usada. Sin embargo, en muestras no muy grandes dichas pruebas producen diferentes resultados. Por lo general, se recomienda el uso de la razón de verosimilitudes.

5.3. R y R^2

Cuando hablamos de regresión lineal, el coeficiente de correlación, r y el de determinación, R^2, son medidas útiles para saber cómo de bien se ajusta el modelo a los datos.

En regresion logística podemos calcular una medida análoga al R^2. Esta medida se conoce como *pseudo* R^2.

Son varias las medidas que se han propuesto para asemejarse al R^2, pero en esta sección, vamos a explicar la de *MacFadden* (1973),

$$R^2_{MacFadden} = \left| 1 - \frac{ln(V)}{ln(V_0)} \right| \tag{5.5}$$

Mediría cuánto del error del ajuste disminuye al incluir las variables predictoras al modelo proporcionando una medición de la significación real del modelo. Además,

- Se trata de la correlación parcial entre la variable respuesta y cada una de las predictoras.
- Puede variar entre 0 y 1.
- Un valor cercano a 1 significa que al crecer la variable predictora, lo hace la probabilidad de que el evento ocurra.
- Un valor cercano a 0 implica que si la variable predictora decrece, la probabilidad de que el resultado ocurra disminuye.
- Si una variable tiene un valor pequeño de $R^2_{MacFadden}$ contribuye al modelo sólo en una pequeña cantidad.

Si volvemos a los datos de `Default`, tenemos,

$$R^2_{MacFadden} = \left| 1 - \frac{-798{,}225}{-1460{,}325} \right| = 0{,}453$$

En este caso, el modelo explica un 45,3 % lo que se conisdera un ajuste moderado.

5.4. Examen de Residuos

Aunque el modelo propuesto represente suficientemente bien los datos, hay una serie de cuestiones que investigar. Podría ocurrir que no se cumpla el supuesto de linealidad entre el *logit* de la probabilidad del suceso de interés y la variable independiente, que la presencia de algunas observaciones extremas en el conjunto de datos que se está manejando perturbe la calidad del ajuste, etc.

Existen numerosos métodos para comprobar el cumplimiento del supuesto de linealidad entre el *logit* de la probabilidad del suceso de interés y las variables independientes. Uno de ellos es el *análisis de residuos* en la regresión logística. Como sabemos, el *residuo* se trata de una medida que expresa la diferencia entre

la respuestas observadas y las predichas por el modelo. Los residuos del modelo (que a veces se denominan *residuos de Pearson*) se definen como,

$$r_i = \frac{(y_i - \hat{p}_i)}{\sqrt{p_i(1-p_i)}} \tag{5.6}$$

Otros residual de interés es la raíz cuadrada de la contribución de cada observación a la lejanía, para el caso en el que $y_i = 0$, toma la forma siguiente:

$$d_{i'} = -\sqrt{-2log(1-p_i)} \tag{5.7}$$

y, para el que caso en que $y_i = 1$, esta otra,

$$d_i = -\sqrt{-2log(p_i)} \tag{5.8}$$

6. Referencias

1 Silva Ayçaguer, L. C., & Barroso Utra, I. M. (2004). *Cuadernos de Estadística. Regresión Logística*. Editorial Hespérides, S.L.

2 James, G., Witten, D., Hastie, T., & Tibshirani, R. (June 21, 2023). *An Introduction to Statistical Learning with Applications in R* (2nd ed., pp. 129-137).

3 Montgomery, E., Vining, D., & Peck (2006). *Introducción Al Análisis de Regresión Lineal* (3rd ed.). Ejemplo 2.1. México: Cecsa.

4 Estimación de máxima verosimilitud en regresión logística (April 26, 2021). Recuperado de: https://arunaddagatla.medium.com/maximum-likelihood-estimation-in-logistic-regression-f86ff1627b67#:~:text=The%20Maximum%20Likelihood%20Estimation%20

5 Ferre Jaén, M. E. (2019). *FEIR 45: Regresión logística*.

6 Regresión logística simple y múltiple (2016). Recuperado de: https://rpubs.com/Joaquin_AR/229736

7 Kaggle. (2018). *Predicción de diabéticos mediante regresión logística*. Data Set. Recuperado de: https://www.kaggle.com/datasets/kandij/diabetes-dataset/data

7. Anexo

```
##############################################################################
############################## R: TRABAJO ####################################
##############################################################################


##############################################################################
############################## EJEMPLO: DISPERSION DE LOS DATOS vs REGRESION LINEAL
##############################################################################
########################################################################## Y CUANTITATIVA
## https://fhernanb.github.io/libro_regresion/rls.html
file <- "https://raw.githubusercontent.com/fhernanb/datos/master/propelente"
datos <- read.table(file=file, header=TRUE)
head(datos) # shows the first 6 rows

library(ggplot2)
ggplot(datos, aes(x=Edad, y=Resistencia)) +
  geom_point() + theme_light()

mod1 <- lm(Resistencia ~ Edad, data=datos);

# GRAFICA: REGRESION LINEAL (Figura 2.2)
ggplot(datos, aes(x=Edad, y=Resistencia)) +
  geom_point(color='#454545', stroke = 1) +
  labs(title = "Regresion lineal por minimos cuadrados") +
  geom_smooth(method='lm', formula=y~x, se=FALSE, col='mediumslateblue', size=1) +
  theme_light() +
  theme(panel.grid = element_blank(),
        panel.background = element_rect(fill = "white", color = "black"),  # Fondo blanco
        panel.border = element_rect(color = "black", fill = NA, size = 0.5))


##############################################################################
########################################################################## Y CUALITATIVA
library(tidyverse)
library(ISLR)
datos <- Default

# Se recodifican los niveles No, Yes a 1 y 0
datos <- datos %>%
  select(default, balance) %>%
  mutate(default = recode(default,
                          "No"  = 0,
                          "Yes" = 1))
head(datos)


# GRAFICA:  REGRESION LINEAL (Figura 2.2)
ggplot(data = datos, aes(x = balance, y = default)) +
  geom_point(aes(color = as.factor(default)), shape = 1) +
  geom_smooth(method = "lm", color = "mediumslateblue", se = FALSE, size=1) +
  scale_color_manual(values = c("#454545", "#454545")) +
  theme_bw()  +
  labs(title = "Regresion lineal por minimos cuadrados",
       y = "Probabilidad default") +
  theme(panel.grid = element_blank(),
        panel.background = element_rect(fill = "white", color = "white"),
```

```
        legend.position = "none")



# GRAFICA: FUNCION logit POR DEFECTO (Figura 3)
logit_function <- function(x) {
  return(log(x / (1 - x)))
}

data <- data.frame(x = seq(0.01, 0.99, 0.01), y = rnorm(99))


ggplot(data, aes(x, y)) +
  stat_function(fun = logit_function, color = "mediumslateblue", size=1) +
  labs(title = "Funci n Logit", y = "y") +
  theme_minimal() +
  theme(panel.grid = element_blank(),
        panel.background = element_rect(fill = "white", color = "black"))

# AJUSTE A UN MODELO DE REGRESION LOGISTICO
modelo_logistico <- glm(default ~ balance, data = datos, family = "binomial")
summary(modelo_logistico)


# GRAFICA: MODELO DE REGRESION LOGISTICO (Figura 3)
ggplot(data = datos, aes(x = balance, y = default)) +
  scale_color_manual(values = c("#454545", "#454545"))+
  geom_point(aes(color = as.factor(default)), shape = 1, alpha = 0.7) +
  stat_function(fun = function(x) {
    predict(modelo_logistico,
            newdata = data.frame(balance = x),
            type = "response")
  }, color = "mediumslateblue", size = 1) +
  theme_minimal() +
  labs(title = "Regresi n log stica",
       y = "Probabilidad default") +
  theme(legend.position = "none",
        panel.grid = element_blank(),
        panel.background = element_rect(fill = "white"))


# Coeficientes Estimados del modelo
modelo_logistico$coefficients
# LejanIa del modelo con la variable independiente
modelo_logistico$deviance
# Lejania del modelo nulo
modelo_logistico$null.deviance
# Razon de verosimilitudes
modelo_logistico$null.deviance-modelo_logistico$deviance
qchisq(0.95,1, lower.tail = T) # test
# R^2 MacFadden
mod_nulo_ln2 <- modelo_logistico$null.deviance/(-2)
mod_ln2 <- modelo_logistico$deviance/(-2)
R_2 = abs(1-(mod_ln2/mod_nulo_ln2)); R_2
```

```
################################################################################
########## EJEMPLO PRACTICO EN R: REGRESION LOGISTICA ##########################
################################################################################
# Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales
# DATOS: Pacientes mujeres de 21 a os o mas que pertenecen a la herencia india
# Pima (subgrupo de nativos americanos)
################################################################################
################################################################################
# LECTURA DE DATOS:
diab <- read.table( "diabetes2.csv", sep = ",", head = TRUE )
head(diab)
attach(diab)
str(diab)

# Variable Dependiente, Y=Outcome
# Y=1 tiene diabetes, Y=0 no tiene diabetes

# Variable Independiente, X=Glucosa

################################################################################
################################################################################
# DIAGRAMA DE DISPERSION:
colores = c()
colores[Outcome == 0] = "#6959CD"
colores[Outcome == 1] = "#FF82AB"

plot(Outcome ~ Glucose, diab, col = colores,
     main = "Diagrama de Dispersion",
     ylab = "Diabetes",
     xlab = "Glucosa", pch = "I")
legend("left", c("Diabetes", "No Diabetes"), pch = 21,
       pt.bg = c("#FF82AB", "#6959CD"))

range(Glucose) # rango de X

################################################################################
################################################################################
# 3. AJUSTE A UN MODELO DE REGRESION LOGISTICO
logmodel <- glm(Outcome ~ Glucose, data = diab, family = binomial( ) )
logmodel
# p(X)= -5.35008 + 0.03787X

################################################################################
################################################################################
# GRAFICO DE UN MODELO DE REGRESION LOGISTICO
# Codificaci n 0,1 de la variable respuesta
Outcome <- as.character(Outcome)
Outcome <- as.numeric(Outcome)

plot(Outcome ~ Glucose, diab, col = colores,
     main = "Diagrama de Dispersion",
     ylab = "Diabetes",
     xlab = "Glucosa", pch = "I")


# type = "response" devuelve la funcion logistica. Si no apareciera, R nos
# nos devuelve la relacion lineal del log-odds y la ecuacion de la recta
curve(predict(logmodel, data.frame(Glucose = x), type = "response"),
      col = 2, lwd = 2, add = TRUE)
```

```
legend("left", c("Diabetes", "No Diabetes"), pch = 21,
       pt.bg = c("#FF82AB", "#6959CD"))

################################################################################
################################################################################
#  3.1 INTERPRETACION DE LOS COEFICIENTES
logmodel$coefficients[2]
# Relacion Positiva entre la glucosa y los diabetes.
# Por cada incremento de la variable glucosa:
exp(logmodel$coefficients[2])

################################################################################
################################################################################
#  3.3 CONVERTIR PROBABILIDAD EN CLASIFICACION
# Vamos a establecer un umbral de 0.5. Si la probabilidad predicha glucosa
# es superior a 0.5, se asiga 1, en caso contrario, 0.

predicciones <- ifelse(test = logmodel$fitted.values > 0.5, yes = 1, no = 0)
matriz_confusion <- table(logmodel$model$Outcome, predicciones,
                          dnn = c("observaciones", "predicciones"))
matriz_confusion

# PRECISION:
# Calcula la suma de la diagonal
diagonal_sum <- sum(diag(matriz_confusion)) # diagonal
total_sum <- sum(matriz_confusion) # total
precision <- diagonal_sum / total_sum; precision
# ERROR:
1-precision

################################################################################
################################################################################
#  3.4 PREDICCIONES DEL MODELO DE REGRESION LOGISTICA
# p(X)= (e^{-5.35008+0.03787*X})/(1+e^{-5.35008+0.03787*X})
# Si una mujer tiene la glucosa a nivel de 168, la clasificariamos:
p_X <- (exp(-5.35+0.037*100))/(1+exp(-5.35+0.037*100))
p_X

################################################################################
################################################################################
#  5. CONSTRUCCION Y EVALUACION DE MODELOS
# 5.1 LEJANIA Y RAZON DE VEROSIMILITUD
summary(logmodel)

dev <- logmodel$deviance
nullDev <- logmodel$null.deviance
RV <- nullDev - dev; RV
qchisq(0.95,1, lower.tail = T)

# 5.2 CONSTRASTE SOBRE LOS PARAMETROS DEL MODELO
summary(logmodel)

# 5.3 R y R^2
mod_nulo_ln2 <- modelo_logistico$null.deviance/(-2)
mod_ln2 <- modelo_logistico$deviance/(-2)
R_2 = abs(1-(mod_ln2/mod_nulo_ln2)); R_2

################################################################################
################################################################################
```