MUBD

Màster Universitari en Enginyeria de Dades Massives (Big Data)

Estadística

laSalle ENG
Universitat Ramon Llull

Índice

Regresión logística

Motivación

- Hay diversas situaciones en que la variable dependiente es éxito o fracaso:
  - Venta (Sí/No) de un producto
  - Moroso (Sí/No) en una entidad financiera
  - Pieza defectuosa (Sí/No) en una cadena de montaje

- Al introducir factores que predigan esta respuesta, no se puede utilizar directamente el modelo lineal general:
  - La respuesta no es Normal. De hecho es dicotómica, sólo puede tomar 2 valores.
  - Al hacer predicciones con posibles valores de las variables predictoras, se podrían obtener estimaciones sin ningún sentido (p.ej: 2.3 , -4.5)
  - Se debe crear una función de conexión (link) que convierta la probabilidad de éxito (entre 0 y 1) en una variable respuesta que tome valores en todos los reales

Regresión logística

Función Link

- La función link es una transformación que se aplica a la probabilidad de éxito (π) para obtener valores que vayan de -∞ a +∞
  - π se mueve entre 0 y 1
  - $\frac{\pi}{1-\pi}$ se mueve entre 0 y +∞. El cociente entre la probabilidad de éxito y la probabilidad de fracaso se denomina ***odd***.
  - $\log\left(\frac{\pi}{1-\pi}\right)$ se mueve entre *-∞* y *+∞.*
- Por tanto, la función link es $\log\left(\frac{\pi}{1-\pi}\right)$
  - La función $\mathrm{f(x)} = \log\left(\frac{x}{1-x}\right)$ se denomina función ***logit.*** No es la única función link, pero sí la más usual e interpretable.
  - Cuando la función *logit* se aplica a una probabilidad, entonces el resultado se denomina ***logodd*** (abreviatura de logaritmo de un odd)

Regresión logística

Modelo

- Por tanto el modelo a ajustar será:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_p \cdot X_p + \varepsilon$$

- Si se desea conocer la probabilidad para un conjunto concreto de variables predictoras y ya se han estimado los coeficientes del modelo, se debe deshacer la transformación:

$$PL = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_p \cdot X_p$$

$$\log\left(\frac{\pi}{1-\pi}\right) = PL \rightarrow \boldsymbol{\frac{\pi}{1-\pi}} = \boldsymbol{odd} = \boldsymbol{e^{PL}} \rightarrow \pi = e^{PL} - \pi \cdot e^{PL} \rightarrow (1 + e^{PL}) \cdot \pi = e^{PL} \rightarrow \boldsymbol{\pi} = \boldsymbol{\frac{e^{PL}}{1 + e^{PL}}}$$

Regresión logística

Forma de la función logit

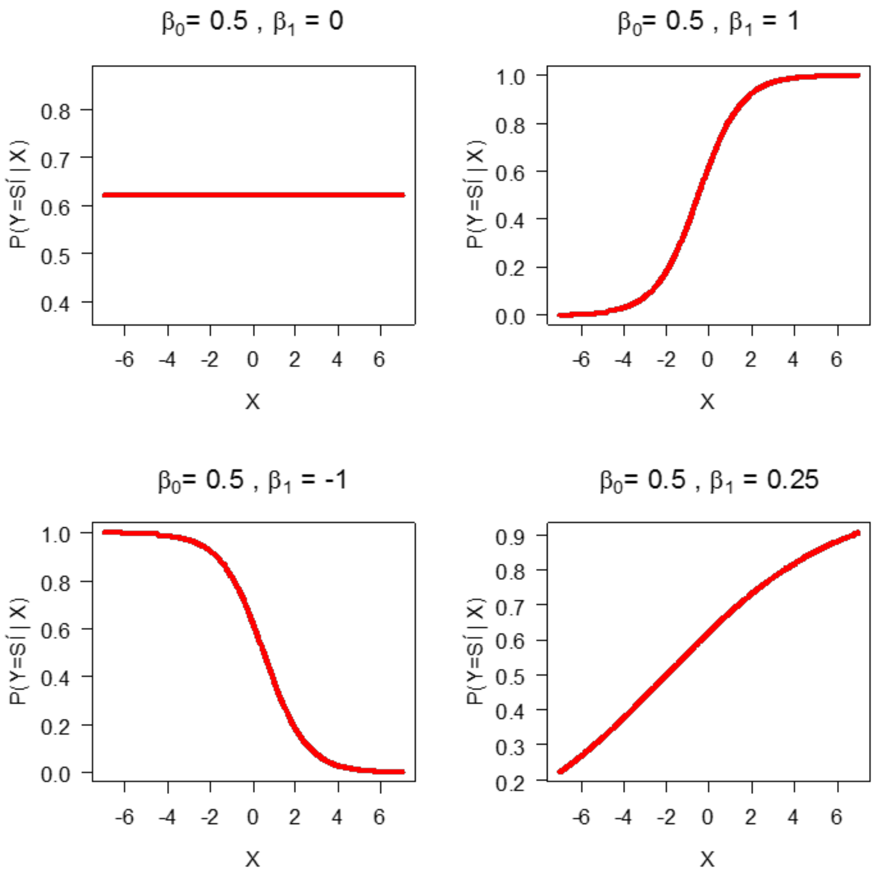
Diferentes formas de la función logit en el caso de una única covariable y con distintos valores de los parámetros.

Sin ‘efecto’ de X

X: factor que favorece el éxito

X: factor que favorece el fracaso

X: factor que favorece levemente el éxito



Regresión logística

Estimación de los parámetros del modelo

- Los parámetros se estiman por el método de **máxima verosimilitud**.
- Este método asume que la variable respuesta sigue una distribución **binomial**.
- Lo que se hace es optimizar el valor de los parámetros del modelo que maximizan la "probabilidad" de haber observado estos datos.
- En R, la función *glm* con el parámetro *family=binomial* realiza el ajuste del modelo

$$\log\left(\frac{\pi}{1-\pi}\right) = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + \cdots + b_p \cdot X_p + e_i$$

- Las *b's* son las estimaciones de las *β's* reales
- **Atención**: Si existe alguna variable categórica con alguna categoría en la que no tiene éxitos o fracasos (problema de las **celdas vacías**), entonces, el modelo no es capaz de estimar el coeficiente para esta categoría.

Interpretación de los coeficientes

Odds ratio (OR)

- El Odds Ratio (OR) se define como el cociente de 2 odds (~riesgo). Ejemplo en tabla 2x2

	Ventas	No ventas	Total
Oferta web	40	20	60
Mailing	20	40	60
Total	60	60	120

$$\left.\begin{aligned} Odd_{web} &= \frac{40}{20} = 2 \\ Odd_{mailing} &= \frac{20}{40} = 0.5 \end{aligned}\right\} \Rightarrow OR = \frac{Odd_{web}}{Odd_{mailing}} = \frac{2}{0.5} = 4$$

- La odds de venta vía web se incrementan por 4 respecto a la venta por mailing
- A través de los coeficientes del modelo, podemos estimar el OR simplemente haciendo la exponencial de los mismos.
- Aunque es parecido, la odd no es idéntica a la probabilidad ($p.ej., P_{web} = \frac{40}{60} = 0.66$)

Interpretación de los coeficientes

Odds ratio (OR)

- El Odds Ratio se define como el cociente de 2 odds (~riesgo)
- Una vez estimado el modelo se puede calcular el OR entre 2 elementos:
  - Odd para el individuo 1: $\exp\{b_0 + b_1 \cdot \mathrm{X}_{11} + \mathrm{b}_2 \cdot \mathrm{X}_{21} + \cdots + \mathrm{b}_\mathrm{p} \cdot \mathrm{X}_{\mathrm{p}1}\}$
  - Odd para el individuo 2: $\exp\{b_0 + b_1 \cdot \mathrm{X}_{12} + \mathrm{b}_2 \cdot \mathrm{X}_{22} + \cdots + \mathrm{b}_\mathrm{p} \cdot X_{p2}\}$
  - Odds ratio:

$$OR = \frac{\exp\{b_0 + b_1 \cdot \mathrm{X}_{11} + \mathrm{b}_2 \cdot \mathrm{X}_{21} + \cdots + \mathrm{b}_\mathrm{p} \cdot \mathrm{X}_{\mathrm{p}1}\}}{\exp\{b_0 + b_1 \cdot \mathrm{X}_{12} + \mathrm{b}_2 \cdot \mathrm{X}_{22} + \cdots + \mathrm{b}_\mathrm{p} \cdot \mathrm{X}_{\mathrm{p}2}\}} =$$

$$= \exp\{\left(b_0 + b_1 \cdot \mathrm{X}_{11} + \mathrm{b}_2 \cdot \mathrm{X}_{21} + \cdots + \mathrm{b}_\mathrm{p} \cdot \mathrm{X}_{\mathrm{p}1}\right) - \left(b_0 + b_1 \cdot \mathrm{X}_{12} + \mathrm{b}_2 \cdot \mathrm{X}_{22} + \cdots + \mathrm{b}_\mathrm{p} \cdot \mathrm{X}_{\mathrm{p}2}\right)\}$$

$$= \exp\{b_1 \cdot (\mathrm{X}_{11} - \mathrm{X}_{12}) + \mathrm{b}_2 \cdot (\mathrm{X}_{21} - X_{22}) + \cdots + \mathrm{b}_\mathrm{p} \cdot (\mathrm{X}_{\mathrm{p}1} - X_{p2})\}$$

- Se puede calcular el OR para el cambio unitario en una variable concreta manteniendo el resto constante. P.ej, si $\mathrm{X}_{11} - \mathrm{X}_{12} = 1$ y $\mathrm{X}_{\mathrm{k}1} - \mathrm{X}_{\mathrm{k}2} = 0$ para $k \neq 1$.
  Nota: Las variables categóricas se codifican con 0’s (referencia) y 1’s (resto de categorías)
- Un OR = 2 en el individuo A respecto al B representa que A tiene el doble de opciones de tener respuesta positiva que B

Regresión logística

Interpretación de los coeficientes

```
Call:
glm(formula=y~status+duration+credit.hist+purpose+credit.amo+savings+installment+debtors+age+installment2+foreign,family=binomial,data=datos)

Coefficients:
                                                   Estimate Std. Error z value Pr(>|z|)
(Intercept)                                       1.006e+00  1.176e+00   0.855 0.392390
status>= 200 DM                                   1.382e+00  4.272e-01   3.236 0.001213 **
status0 <=  and < 200 DM                          6.522e-01  2.533e-01   2.575 0.010035 *
statusno checking account                         2.086e+00  2.769e-01   7.535 4.90e-14 ***
duration                                         -2.832e-02  1.093e-02  -2.591 0.009556 **
credit.histcritical account                       1.756e+00  4.895e-01   3.587 0.000335 ***
credit.histdelay in paying off in the past        8.378e-01  5.257e-01   1.594 0.111038
credit.histexisting credits paid back duly till now      1.060e+00  4.559e-01   2.324 0.020102 *
credit.histno credits taken/ all credits paid back duly  6.783e-01  6.391e-01   1.061 0.288562
purposecar (new)                                 -1.067e+00  3.943e-01  -2.707 0.006798 **
…
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 852.29  on 686  degrees of freedom
Residual deviance: 620.05  on 657  degrees of freedom
AIC: 680.05
```

Tener mas de 200 marcos en la cuenta incrementa en 4 ($e^{1.38}$) las probabilidades de pagar el crédito respecto a la referencia (estar en negativo)

El error esperado en la estimación del coeficiente es 0.42

Esta categoría es significativa ($p < 0.05$)

Cada mes de más en el número total de meses del crédito baja la probabilidad de pagar el crédito un 3% ($e^{-2.83}$)

Selección de variables

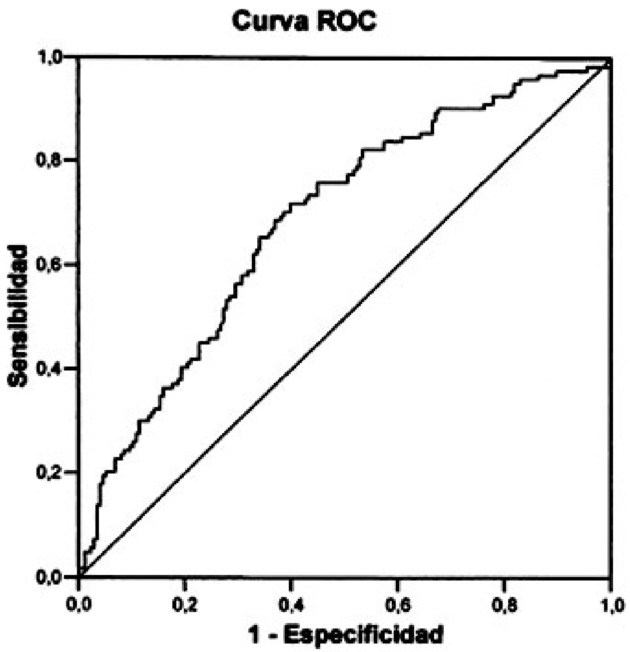
Métodos automáticos vs Métodos Manuales

- Los métodos vistos en el modelo lineal son aplicables en el caso de la regresión logística
- Hay estudios que demuestran que los métodos automáticos inflan la capacidad predictiva y son menos replicables que los métodos manuales
- Sin embargo, los métodos automáticos basados en *stepwise* muestran resultados aceptables con relativamente poco esfuerzo
- En R, la instrucción *step* nos realiza la selección automática basándose en el criterio del AIC ya visto

Capacidad predictiva

Curva ROC y AUC

- Definiciones:
  - Sensibilidad: Proporción de individuos con probabilidades predichas más altas que un punto de corte fijado entre aquellos con respuesta positiva. (P. ej. proporción de personas con una probabilidad de venda predicha mayor que 0.5 entre los que realmente se ha vendido)
  - Especificidad: Proporción de individuos con probabilidades más bajas que el punto de corte dentro de aquellos con respuesta negativa (P. ej. Proporción de personas con una probabilidad de venda predicha inferior a 0.5 entre los que NO se han vendido)



- La curva ROC representa, para cada punto de corte de la probabilidad predicha, la sensibilidad en función del complementario de la especificidad (1-especifidad).
- Cuanto más *abombada* (menos plana) sea la curva, mejor capacidad predictiva

Capacidad predictiva

Ejemplo

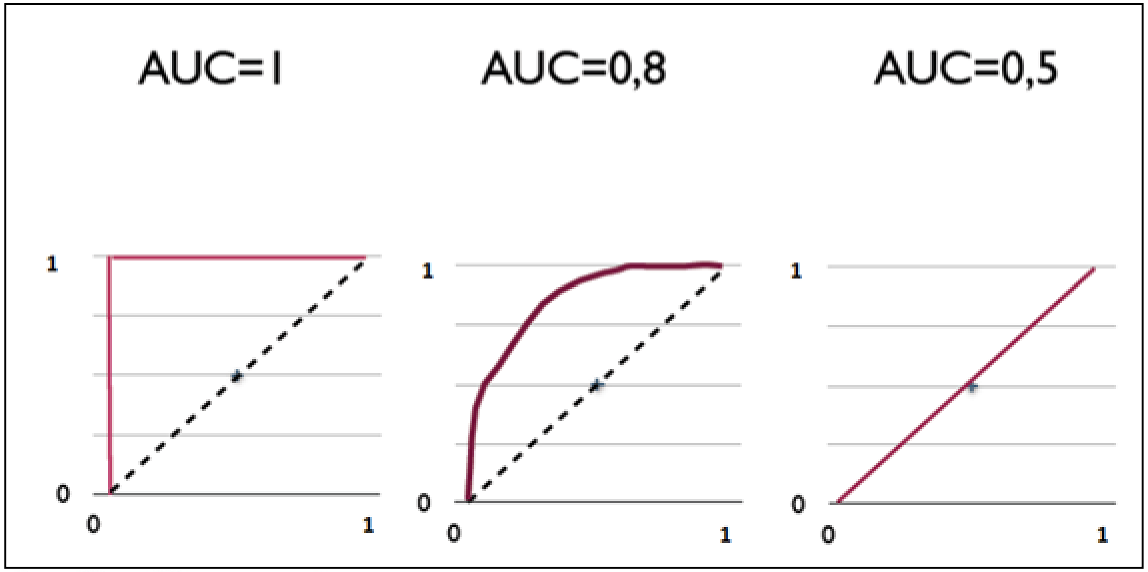
Probabilidad Predicha	Respuesta Real
0.08	No venta
0.27	No venta
0.34	Venta
0.36	No venta
0.44	No venta
0.52	No venta
0.53	Venta
0.54	No venta
0.80	Venta
0.96	Venta

- **Punto de corte π = 0.4**
  - Dentro de los que tienen respuesta positiva (Respuesta = Venta), hay 3 con probabilidades predichas superiores a 0.4 y 1 por debajo. Por tanto la **Sensibilidad** es del 75%
  - Dentro de los que tienen respuesta negativa, hay 3 con probabilidades predichas inferiores a 0.4 y 3 por encima. Por tanto la **Especificidad** es del 50%
  - El punto a representar en la curva ROC seria (0.50, 0.75)
- **Punto de corte π = 0.5**
  - Dentro de los que tienen respuesta positiva (Respuesta = Venta), hay 3 con probabilidades predichas superiores a 0.5 y 1 por debajo. Por tanto la **Sensibilidad** es del 75%
  - Dentro de los que tienen respuesta negativa, hay 4 con probabilidades predichas inferiores a 0.5 y 2 por encima. Por tanto la **Especificidad** es del 66.7%
  - El punto a representar en la curva ROC seria (0.33, 0.75)

Capacidad predictiva

Curva ROC y AUC

- El AUC (Area Under Curve) representa el área que queda debajo de la curva ROC. Cuánto mayor sea esta área, mayor capacidad predictiva del modelo.
- Una interpretación de la curva ROC es la proporción de parejas de observaciones (respuesta positiva – respuesta negativa) donde el individuo con respuesta positiva tiene una probabilidad predicha mayor.
- Un AUC = 1 implica una discriminación perfecta y un AUC = 0.5 implica discriminación al azar



Interpretación AUC:

[0.5, 0.6): Discriminación mala

[0.6, 0.75): Discriminación regular

[0.75, 0.9): Discriminación buena

[0.9, 1): Discriminación muy buena

Validación

Bondad del ajuste

- El estadístico de Hosmer-Lemeshow nos da una medición de la bondad del ajuste de modelo comparando los efectivos esperados y los observados en cada tramo de probabilidad
  - Divide la probabilidades predichas en *k* (usualmente 10) intervalos. P.ej, de 0 a 1 en intervalos de 0.1
  - Se obtiene por una lado, el número de respuestas positivas en cada intervalo (**obs: efectivos observados**)
  - Se obtiene por otro lado, la suma de probabilidades predichas para cada intervalo (**esp: efectivos esperados**)
  - Con el estadístico chi-cuadrado se evalúa si el modelo es válido:

$$\chi^2 = \sum \left(\frac{(esp - obs)^2}{obs} \right)$$

  - Si el p-valor de la prueba es inferior a 0.05, no se valida el modelo porque los efectivos esperados y observados en cada tramo discrepan demasiado.
  - **El inconveniente de esta prueba es que para tamaños de muestra grandes, siempre se rechaza la validez. En estos caso es mejor comparar visualmente los intervalos**

MUBD

Màster Universitari en Enginyeria de Dades Massives (Big Data)

Estadística