Regresión logística

Mathias Bourel

IMERL - Facultad de Ingeniería, Universidad de la República, Uruguay

October 10, 2019

Plan

Introducción

Supongamos que queremos modelar una variable Y categórica, binaria, ($Y \sim B(1, p)$) por ejemplo:

- Presencia/ausencia de una determinada especie
- Especie1/especie2
- Enfermo/ no enfermo
- Spam/no spam
- Quiebra/no-quiebra

Si quisieramos usar la regresión lineal

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$\mathbb{E}(Y|X = x_i) = \beta_0 + \beta_1 x_i$$

$$\mathbb{E}(Y|X = x_i) = \underbrace{\mathbb{P}(Y = 1|X = x_i)}_{p_i} \times 1 + \mathbb{P}(Y = 0|X = x_i) \times 0 = p_i$$

y por lo tanto

$$p_i = \beta_0 + \beta_1 x_i$$

O sea la predicción hecha por el modelo estima la probabilidad de que el individuo x_i pertenezca a la población 1. Inconveniente: $p_i \in [0, 1]$....No parece apropiado...

Definición de los Modelos Lineales Generalizados

Volviendo al caso general, si llamamos $\mu = \mathbb{E}(Y|X)$, y consideramos una función g monótona y diferenciable entonces

$$g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = \nu$$

Un modelo lineal generalizado está dado por

$$g(\mu) = X'\beta = \nu$$

donde ν es un predictor lineal. Todo GLM tiene:

- una componente aleatoria: variable de respuesta Y, representada por μ.
- una componente sistemática: combinación lineal de las variables explicativas (independientes, predictoras).
- Función link o de enlace: relaciona las dos componentes anteriores.

Plan

Ejemplo, regresión logística

```
>library(ISLR)
>attach(Default)
>data=Default
>data
>head(data,n=4)
 default student   balance   income
1       No      No  729.5265 44361.63
2       No     Yes  817.1804 12106.13
3       No      No 1073.5492 31767.14
4       No      No  529.2506 35704.49
```

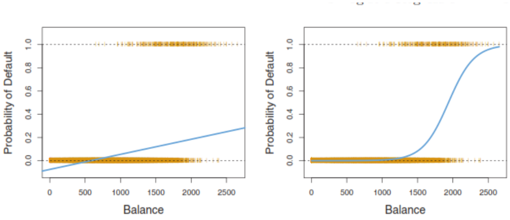

Figure: Ejemplo: Probabilidad de "default" en la tarjeta de crédito en función del balance mensual en la tarjeta (Cap. 4 de [2])

Claramente la recta de regresión lineal no se ajusta bien a los datos por lo cual preferimos una sigmoide.

Modelo Logístico

Volvemos al modelo de regresión logística binaria. Vamos a querer que

$$p(x_i) = p_i = F(\beta_0 + \beta_1 x_i)$$

donde F es una función de distribución para que el modelo proporcione directamente la probabilidad de pertenecer a cada uno de los grupos.
Por lo que una función link adecuada para modelar este tipo de variables es la función *logit* quedando el modelo de regresión logística con la siguiente forma: si $p = \mathbb{P}(Y = 1|X)$ entonces el modelo logístico múltiple es:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Entonces

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}}$$

En este caso $F(t) = \frac{1}{1+e^{-t}}$ y se llama función de distribución logística.

Estimación de los parámetros en regresión logística

A partir de n observaciones y suponiendo que

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \quad i = 1, \ldots, n$$

se buscan los paramétros $\theta = (\beta_0, \beta_1, \ldots, \beta_d)$ que maximicen el logaritmo de la función de verosimilitud L:

$$\ln(L(y,\theta)) = \ln\left(\prod_{i=1}^{n} f(y_i,\theta)\right) = \sum_{i=1}^{n} \ln f(y_i,\theta)$$

En el caso de la regresión lógistica binaria, y suponiendo que el modelo es $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \mathbf{x}$, la función de verosimilitud en el caso que no haya datos repetidos se calcula como:

$$L(y,\theta) = L(y,\beta_0,\beta_1) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

y

$$\ln(L(y,\beta_0,\beta_1)) = \sum_{i=1}^{n}(y_i \ln(p_i) + (1-y_i)\ln(1-p_i))$$

Acordarse que $p_i = \frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)} \quad \forall\, i = 1, \ldots, n.$

Estimación de los parámetros

Se prueba que se encuentra un único vector β que anula a todas las derivadas parciales de $L(y, \theta)$ simultaneamente. Ese β resulta ser un óptimo del problema de maximización.

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \underset{\beta_0, \beta_1}{\operatorname{Argmax}} \ln(L(y, \beta_0, \beta_1))$$

En la práctica este estimador se calcula usando métodos iterativos (método de Newton-Raphson o Fisher scoring).

Test sobre significancia modelo

Se puede usar la desvianza nula y la desvianza residual para testear la significancia del modelo:

$$\begin{array}{ll} (H_0): & \beta_j = 0 \, \forall \, j = 1, \ldots, p \\ (H_1): & \exists \beta_j \neq 0 \end{array}$$

Bajo la hipótesis nula (H_0) el logaritmo del cociente de las verosimilitudes $-2\ln\left(\frac{L_{\text{nulo}}}{L_{\text{completo}}}\right) \sim \chi^2_p$

$-2\ln\left(\frac{L_{\text{nulo}}}{L_{\text{completo}}}\right) = -2\ln(L_{\text{nulo}}) + 2\ln(L_{\text{completo}}) = \text{Null deviance} - \text{Residual deviance}$

```
>modelo.null=glm(default~1,data=Default, family='binomial')
>modelo3=glm(default~.,data=Default, family='binomial')
> anova(modelo.null,modelo3,test='Chisq')
Analysis of Deviance Table

Model 1: default ~ 1
Model 2: default ~ student + balance + income
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      9999     2920.7
2      9996     1571.5  3   1349.1 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Como el p-valor es menor que 0.05 hay suficiente evidencia para rechazar la hipotesis nula
Teniendo en cuenta que los grados de libertad son $p = 3$, otra posibilidad es hacer:

```
> chi= modelo.null$deviance - modelo3$deviance
> pchisq(chi, df=3,lower.tail=F)
```

y llegamos a la misma conclusión.

Modelos Anidados

Esto se puede extender a la comparación de dos modelos anidados

$$\begin{cases} (H_0): & \beta = (\beta_0, \beta_1, \ldots, \beta_q) \\ (H_1): & \beta = (\beta_0, \beta_1, \ldots, \beta_p) \end{cases}$$

con $q < p < n$. La idea es que si la hipótesis nula es cierta entonces las verosimilitudes deben ser muy cercanos en valor y por lo tanto la diferencia entre los logaritmos chica. Usamos las diferencias entre las desvianzas

$$D_0 - D_1 = 2\left(\ln(L(\widehat{\beta}_p, y)) - \ln(L(\widehat{\beta}_q, y))\right)$$

Si los dos modelos describen bien los datos entonces $D_0 \sim \chi^2_{n-(q+1)}$ y $D_1 \sim \chi^2_{n-(p+1)}$ por lo tanto $D_0 - D_1 \sim \chi^2_{p-q}$ y rechazamos la hipotesis nula si $D_0 - D_1 > \chi^2_{p-q}$.

Modelos Anidados

Aca vemos la prueba chi-cuadrado a medida que vamos añadiendo las variables.

```
> anova(modelo3,test='Chisq')
Analysis of Deviance Table

Model: binomial, link: logit

Response: default

Terms added sequentially (first to last)


        Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     9999     2920.7
student  1    11.97      9998     2908.7 0.0005416 ***
balance  1  1337.00      9997     1571.7 < 2.2e-16 ***
income   1     0.14      9996     1571.5 0.7115139
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
>
```

Criterio Akaike

Al igual que vimos para los modelos lineales (ML), podemos usar el criterio de Akaike (AIC) para comparar modelos con distinto número de parámetros.

$$AIC = -2\ln(L(y, \widehat{\beta}_p)) + 2(p+1)$$

Recordemos que cuanto menor el valor del AIC, mejor es el ajuste.
El número AIC por sí solo no nos dice nada, lo que nos interesa es la diferencia de AIC entre diferentes modelos.

Criterio posible:
diferencias de 0 a 2: modelos similares
diferencias de 4 a 7: es mejor el modelo con menor AIC
diferencias > 10: es mucho mejor el modelo con menor AIC

Si tenemos muchas variables podemos usar selección de variables con los métodos stepwise (paso a paso), forward (hacia adelante) o backward (hacia atrás) tomando como criterio de selección en cada paso el valor del AIC.

Comparación de modelos

```
> anova(modelo,modelo3,test='Chisq')
Analysis of Deviance Table

Model 1: default ~ balance
Model 2: default ~ student + balance + income
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      9998     1596.5
2      9996     1571.5  2   24.907 3.904e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> AIC(modelo)-AIC(modelo3)
[1] 20.90686
```

Hay evidencia por el test chi-cuadrado que el modelo 3 es mejor que el modelo1 y por otro lado también con el criterio del AIC

Estimación e interpretación de los coeficientes

- La estimación de los coeficientes de la regresión hecha por el método de máxima verosimilitud se aplica también para cualquier GLM. Idem la comparación de modelos.
- Supongamos que $\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \mathbf{x}$ donde $p = p(\mathbf{x}) = \mathbb{P}(Y = 1|X = \mathbf{x})$
  - Odds($\mathbf{x}$) $= \frac{p(\mathbf{x})}{1-p(\mathbf{x})}$ indica cuantas veces es más probable que ocurra $Y = 1$ respecto a que no ocurra (o sea ocurra $Y = 0$).
  - Odds-ratio: OR($\mathbf{x}$)$= \frac{Odds(\mathbf{x}+1)}{Odds(\mathbf{x})}$ (como cambia la respuesta de interés al aumentar una unidad). En este caso una cuenta inmediata muestra que $\beta_1 = \ln(OR)$ y por lo tanto
  $$e^{\beta_1} = OR$$
  Entonces si X augmenta de k unidades se tiene que $OR = e^{k\beta_1}$.
  - Odds-ratio($\mathbf{x_i}, \mathbf{x_j}$)$= \frac{Odds(\mathbf{x_i})}{Odds(\mathbf{x_j})} = e^{\beta_1(\mathbf{x_j}-\mathbf{x_i})}$ y en general es igual a $e^{\beta'(\mathbf{x_j}-\mathbf{x_i})}$

Por ejemplo si:

- la probabilidad de tener cancer de pulmon para un fumador es $\mathbb{P}(Y = 1|X = \text{fumador}) = 0.01$ por lo que $\mathbb{P}(Y = 0|X = \text{fumador}) = 0.99$ y odds(X =fumador)=1/99.
- la probabilidad de tener cancer de pulmon para un no fumador es $\mathbb{P}(Y = 1|X = \text{no fumador}) = 10^{-4}$
- OR(fumador, no fumador)$= \frac{1/99}{1/9999} = 101$ y hay 101 veces más chance de tener cancer de púlmon para un fumador que para un no fumador.

Estimación e interpretación de los coeficientes

- Recordemos que en la regresión lineal el coeficiente β_j asociado a una determinada variable X_j indicaba el cambio en la variable Y al aumentar una unidad de la variable X_j manteniendo las demás fijas.
- Aquí, lo que nos dice este coeficiente es el cambio en el log(p/1-p) al aumentar en una unidad la X_j.
- Una forma de interpretar los coeficientes β_j de forma genérica es: si son positivos, entonces al aumentar X_j aumenta la probabilidad de ocurrencia de default, si son negativos, al aumentar X_j, esta probabilidad disminuye.

Tabla comparativa LM y GLM

	LM	GLM
Parámetros	$\beta_0, \beta_1, \ldots, \beta_p$	$\beta_0, \beta_1, \ldots, \beta_p$
Estimación	Mínimos Cuadrados	Máxima Verosimilitud
Ajuste	R^2	Desvianza
Comparación modelos	AIC, F	AIC, Desvianza
Supuestos	Residuos normales +Gauss	Y familia exponencial

Seguimos con el ejemplo

```
> modelo=glm(default~balance,family=binomial,data)
> summary(modelo)
Call:
glm(formula = default ~ balance, family = binomial, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2697  -0.1465  -0.0589  -0.0221   3.7589

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
balance      5.499e-03  2.204e-04   24.95   <2e-16 ***
```

El test de hipótesis que aparece es el Test de Wald con estadístico $\frac{(\widehat{\beta_j}-b_j)^2}{\mathrm{Var}(\widehat{\beta_j})} \to \chi^2(1)$ bajo $(H_0):\ \beta_j = 0$. En lo que nos proporciona R, tenemos $\frac{\widehat{\beta_j}}{\mathrm{s.e}(\widehat{\beta_j})} \to \mathcal{N}(0,1)$

- $\widehat{\beta}_1 = 0.0055 \Rightarrow$ incremento en balance implica incremento en default.
- El estadistico $z = \widehat{\beta}_1 / s.e(\widehat{\beta}_1)$ juega el mismo papel que el estadístico t de la regresión lineal por lo que un valor importante de z implica rechazar la hipotesis nula $(H_0) : \beta_1 = 0$. Esta hipótesis nula implica que $p(X) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$ y no depende del valor de X (en este caso balance).
- Claramente hay una relación entre balance y default.
- Predicción:

```
predict(modelo, data.frame(balance=c(1000,2000)), type='response')
0.005752145 0.585769370
```

  Esto es que $\widehat{p}(1000) = \frac{e^{-10.63+0.0055\times 1000}}{1+e^{-10.63+0.0055\times 1000}} = 0.00575, \qquad \widehat{p}(2000) = 0.586$
- Ajuste:

```
>1-pchisq(modelo$deviance,9998)
```

  Obtenemos el valor 1 entonces el modelo se ajusta bien a los datos.
- Si queremos calcular el riesgo de default al aumentar el balance en 100 dólares se tiene que $e^{100\times 0.0055} = 1.73$ y el riesgo aumenta aproximadamente 2 veces.

Usando la variable categorica "student"

```
>modelo2=glm(default~student,family=binomial,data)
>summary(modelo2)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.50413    0.07071  -49.55  < 2e-16 ***
studentYes   0.40489    0.11502    3.52 0.000431 ***


>predict(modelo2, data.frame(student=c("Yes","No")), type='response')
0.04313859 0.02919501
```

$$\mathbb{P}(\text{default=Yes}|\text{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431$$

$$\mathbb{P}(\text{default=Yes}|\text{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292$$

Usando todas las variables:

```
> modelo3=glm(default~.,family=binomial,data)
> summary(modelo3)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
```

```
>predict(modelo3, data.frame( student="Yes",balance=1500,income=40000), type='response')

0.05788194
```

$$\widehat{p}(X) = \frac{e^{-10.869+0.00574\times1500+0.003\times40-0.6468\times1}}{1+e^{-10.869+0.00574\times1500+0.003\times40-0.6468\times1}} = 0.058$$

Si Student=No: $\widehat{p}(X) = \frac{e^{-10.869+0.00574\times1500+0.003\times40-0.6468\times0}}{1+e^{-10.869+0.00574\times1500+0.003\times40-0.6468\times0}} = 0.105$

Las predicciones se hacen generalmente con la regla del *máximo a posteriori*, es decir predecimos el valor de Y por la modalidad k que maximiza la probabilidad $\mathbb{P}(Y = k|X = \mathbf{x})$.
En presencia de dos clases, podríamos pensar que si la probabilidad de estar en una clase es mayor que $1/2$, entonces esa debe ser la clase asignada a $\mathbf{x}$. Pero esta elección de $1/2$ es totalmente arbitraria. Se podría definir la regla de asignación con umbral s como

$$y_s^* = \begin{cases} 1 & \text{si } \mathbb{P}(Y = 1|X = \mathbf{x}) \geq s \\ 0 & \text{si no} \end{cases}$$

¿Qué tan performante es el modelo? ¿Cuánto se equivoca?

```
>pred=predict(modelo3,data.frame=data,type="response")
>contrasts(default)
>prediccion=rep("No",10000)
>prediccion[pred>.5]="Yes"
>table(prediccion,default)

> table(prediccion,default)
          default
prediccion   No  Yes
       No  9627  228
       Yes   40  105
> mean(prediccion==default)
[1] 0.9732
```

Es satisfactorio??....

Modelo logístico multiclasses

Supongamos ahora que Y tiene K modalidades con $K > 2$. Sea $\pi_k = \mathbb{P}(Y = k|X = \mathbf{x})$. Nos fijamos una modalidad de referencia, generalmente la última, K, y hacemos $K - 1$ regresiones logísticas de $\pi_k(\mathbf{x})$ vs $\pi_K(\mathbf{x})$:

$$\ln\left(\frac{\pi_k(\mathbf{x})}{\pi_K(\mathbf{x})}\right) = \beta_{0k} + \sum_{j=1}^{p} \beta_{jk} x_j \quad \forall\, 1 \leq k \leq K - 1$$

La clasificación se hace asignando a $\mathbf{x}$ la clase con la máxima probabilidad a posteriori, es decir, calculamos las K probabilidades a posteriori siguiente:

$$\mathbb{P}(Y = k|X = \mathbf{x}) = P(Y = K|X = x)e^{\beta_{0k} + \sum_{j=1}^{p} \beta_{jk} x_j} \quad \forall\, k \in \{1, \dots, K - 1\}$$

$$\mathbb{P}(Y = K|X = \mathbf{x}) = 1 - \sum_{k=1}^{K-1} \mathbb{P}(Y = k|X = \mathbf{x}) = 1 - \sum_{k=1}^{K-1} P(Y = K|X = \mathbf{x})e^{\beta_{0k} + \sum_{j=1}^{p} \beta_{jk} x_j}$$

$$\mathbb{P}(Y = K|X = \mathbf{x}) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\beta_{0k} + \sum_{j=1}^{p} \beta_{jk} x_j}}$$

y clasificamos $\mathbf{x}$ en aquella clase k que hace máxima $\mathbb{P}(Y = k|X = \mathbf{x})$.

Referencias

1. Mathias Bourel, Carolina Crisci. Notas del curso Estadística Avanzada y Aplicaciones, MAREN, CURE Rocha, 2014.
2. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, 2013.