Tema 3. Minería Estadística de datos. Regresión logística

Beatriz Coronado Sanz

28 de enero de 2019

1. Ejemplo de incendios

Vamos a realizar diferentes modelos de regresión logística sobre el fichero *incendios.csv*, y usaremos el contraste fundamental de hipótesis con $\alpha = 0{,}05$ para determinar el ajuste de los modelos. Después realizaremos una predicción para un municipio ficticio y calcularemos las odds ratio de las variables del modelo.

Lo primero que hacemos es leer los datos del fichero. Los 6 primeros datos son:

```
  maquin_d  gan_for   paro roadmu_d frag7x7 prestur Y
1  473.199   0.2298 0.1926  230.252  0.0171       0 1
2  432.876 113.1860 0.1533  331.600  0.0296       1 1
3  143.164 103.8490 0.1790  295.636  0.0244       0 0
4  502.235  91.1550 0.1377  273.190  0.0244       0 1
5  139.374  81.8590 0.2002  123.163  0.0348       0 0
6  738.364 236.3120 0.1457   52.500  0.0050       0 0
```

Observamos que tenemos 7 variables diferentes. La variable Y indica la alta/baja incidencia al fuego en un determinado municipio. Queremos encontrar una asociación entre el valor de esta variable y el resto.

Veamos un resumen descriptivo de las variables incluidas en el fichero:

```
    maquin_d        gan_for            paro           roadmu_d    
 Min.   : 32.0   Min.   :  0.12   Min.   :0.0204   Min.   : 52.5  
 1st Qu.:219.3   1st Qu.: 58.61   1st Qu.:0.1158   1st Qu.: 95.6  
 Median :357.3   Median :106.27   Median :0.1461   Median :192.0  
 Mean   :358.2   Mean   :109.12   Mean   :0.1464   Mean   :211.6  
 3rd Qu.:496.8   3rd Qu.:155.65   3rd Qu.:0.1798   3rd Qu.:302.3  
 Max.   :956.4   Max.   :299.39   Max.   :0.2756   Max.   :662.3  
    frag7x7           prestur             Y         
 Min.   :0.00500   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:0.02117   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :0.03280   Median :1.0000   Median :1.0000  
 Mean   :0.03308   Mean   :0.5733   Mean   :0.6767  
 3rd Qu.:0.04425   3rd Qu.:1.0000   3rd Qu.:1.0000  
 Max.   :0.07600   Max.   :1.0000   Max.   :1.0000  
```

Lo primero que queremos hacer es indicar que las variables Y y *prestur* son dicotómicas, es decir, que solo admiten los valores 1 o 0. Mostramos a continuaciń el número de elementos para cada uno de los valores de la variable Y:

```
  0   1
 97 203
```

a) Ajuste de un modelo de regresión logística múltiple para relacionar la variable Y con el resto de variables.

$$\ln\left(\frac{\pi(\underline{x})}{1-\pi(\underline{x})}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

En R obtenemos:

```
Call:
glm(formula = Y ~ ., family = binomial, data = datos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8675  -0.3915   0.1925   0.4725   1.9495

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.583862   0.897859  -3.992 6.56e-05 ***
maquin_d     0.011588   0.001466   7.904 2.70e-15 ***
gan_for     -0.008734   0.002997  -2.914  0.00356 **
paro        -4.970866   3.793522  -1.310  0.19008
roadmu_d     0.003505   0.001472   2.380  0.01730 *
frag7x7     23.577807  12.233963   1.927  0.05395 .
prestur1     1.981914   0.390963   5.069 3.99e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 377.61  on 299  degrees of freedom
Residual deviance: 192.15  on 293  degrees of freedom
AIC: 206.15

Number of Fisher Scoring iterations: 6
```

b) Justificar si el modelo ajustado es adecuado basándose en el contraste fundamental de hipótesis.

El modelo es adecuado porque 4 de 6 variables son aceptadas en el modelo y se rechaza para ellas la hipótesis nula. Por tanto, se rechaza la hipótesis nula para el modelo en general.

c) Interpretacion de los resultados obtenidos en los constrastes individuales de los coeficientes y propuesta de modelo final.

Observamos que para las variables *maquin_d*, *prestur1*, *gan_for* y roadmu_d se rechaza la hipótesis nula puesto que tienen un valor menor que $\alpha = 0{,}05$. Las otras dos variables: *paro* y *frag 7x7* no rechazan la hipótesis nula para un valor de $\alpha = 0{,}05$, por lo que el modelo final sería aquel que se calcule sin contar con estas dos variables.

El modelo sin considerar las variables *paro* y *frag 7x7* sería:

```
Call:
glm(formula = Y ~ maquin_d + gan_for + roadmu_d + prestur, family = binomial,
    data = datos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7714  -0.4329   0.1968   0.4810   2.0237

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.462908   0.610963  -5.668 1.45e-08 ***
maquin_d     0.011663   0.001471   7.931 2.18e-15 ***
gan_for     -0.009044   0.003003  -3.011   0.0026 **
roadmu_d     0.003468   0.001431   2.424   0.0154 *
prestur1     1.890110   0.380115   4.972 6.61e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 377.61  on 299  degrees of freedom
Residual deviance: 197.90  on 295  degrees of freedom
AIC: 207.9

Number of Fisher Scoring iterations: 6
```

d) Predicción razonada de la probabilidad de fuego de un municipio con unos ciertos valores para las variables explicativas que se deben considerar en el estudio.

Para nuestra predicción crearemos un ejemplo donde las variables *paro* y *frag 7x7* son 0 para usar nuestro modelo final y, por tanto, no se tendrán en cuenta para predecir la variable Y.

El ejemplo considerado y su predicción se muestran a continuación:

```
  maquin_d gan_for roadmu_d prestur
1      500     150      100       0

       1
0.795589
```

Observamos que el valor del municipio ficticio es cercano a 1 (0,796) y, por lo tanto, tiene alta incidencia de fuego.

e) Cálculo de un intervalo de confianza al 95 % para las odds ratio de las variables significativas del modelo. Interpretación de los resultados obtenidos para las variables *gan_for* y *prestur*.

Las odds ratio y sus intervalos de confianza de las variables significativas del modelo son:

```
            Odds ratio Lim Inf (2.5%) Lim Sup (97.5%)
(Intercept)  0.0313385     0.00879325      0.09752922
maquin_d     1.0117317     1.00902576      1.01488904
gan_for      0.9909963     0.98496094      0.99668741
roadmu_d     1.0034743     1.00075157      1.00640751
prestur1     6.6200954     3.20939868     14.34788626
```

El odd ratio de *gan_for* es 0,99, lo que quiere decir que por cada 1 % más de densidad de ganado en el municipio el riesgo de fuego desciende un 1 %.

El odd ratio de *prestur* es 6,62, lo que quiere decir que en un municipio con presión turística el riesgo es $6,62$ veces mayor que en uno que no tiene presión turística.

2. Ejemplo de admisiones

Vamos a realizar un modelo de regresión logística sobre el fichero *binary.txt*, que recoge los datos de admisión de ciertos alumnos de EEUU en estudios de postgrado. Comprobaremos el ajuste del modelo para $\alpha = 0{,}05$, luego predeciremos la probabilidad de que un alumno sea admitido según unos datos concretos y, por último, calcularemos las odds ratio de las variables del modelo.

Lo primero que hacemos es leer los datos del fichero. Los primeros 6 datos son:

```
  admit gre  gpa rank
1     0 380 3.61    3
2     1 660 3.67    3
3     1 800 4.00    1
4     1 640 3.19    4
5     0 520 2.93    4
6     1 760 3.00    2
```

Observamos que tenemos 4 variables diferentes. La variable *admit* indica si un alumno ha sido admitido o no. Queremos encontrar una asociación entre el valor de esta variable y el resto.

Veamos un resumen descriptivo de las variables incluidas en el fichero:

```
     admit             gre             gpa             rank
 Min.   :0.0000   Min.   :220.0   Min.   :2.260   Min.   :1.000
 1st Qu.:0.0000   1st Qu.:520.0   1st Qu.:3.130   1st Qu.:2.000
 Median :0.0000   Median :580.0   Median :3.395   Median :2.000
 Mean   :0.3175   Mean   :587.7   Mean   :3.390   Mean   :2.485
```

```
 3rd Qu.:1.0000   3rd Qu.:660.0   3rd Qu.:3.670   3rd Qu.:3.000
 Max.   :1.0000   Max.   :800.0   Max.   :4.000   Max.   :4.000
```

Lo primero que queremos hacer es indicar que la variable *admit* es dicotómica. A continuación, mostramos el número de elementos para cada uno de los valores de la variable *admit*:

```
  0   1
273 127
```

a) Ajuste de un modelo de regresión logística múltiple para relacionar la variable *admit* con el resto de variables. Discusión sobre si el modelo es adecuado o no.

```
Call:
glm(formula = admit ~ ., family = binomial, data = misdatos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5802  -0.8848  -0.6382   1.1575   2.1732

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.449548   1.132846  -3.045  0.00233 **
gre          0.002294   0.001092   2.101  0.03564 *
gpa          0.777014   0.327484   2.373  0.01766 *
rank        -0.560031   0.127137  -4.405 1.06e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 499.98  on 399  degrees of freedom
Residual deviance: 459.44  on 396  degrees of freedom
AIC: 467.44

Number of Fisher Scoring iterations: 4
```

Observamos que todas las variables rechazan la hipótesis nula para $\alpha = 0{,}05$, por lo que nuesro modelo es adecuado.

b) Predicción de la probabilidad de que sea admitido un alumno cuya puntuación en el exámen de ingreso fue 525, su media en el grado es 3,1 y estudió en una institución cuyo rango es 3.

Creando una variable para el alumno que nos preguntan y utilizando la función *predict* llegamos a que la probabilidad de que sea admitido este alumno es:

```
        1
0.1799669
```

Lo que quiere decir que, con un 17 %, el alumno no será admitido en esa universidad.

c) Cálculo de las odds ratios de las variables significativas del modelo e interpretación de los resultados obtenidos.

Las odds ratio y sus intervalos de confianza de las variables del modelo son:

	Odds ratio	Lim Inf (2.5%)	Lim Sup (97.5%)
(Intercept)	0.03175998	0.003309497	0.2835650
gre	1.00229659	1.000171559	1.0044714
gpa	2.17496718	1.152082367	4.1717746
rank	0.57119114	0.442656492	0.7294389

Observamos que la odd ratio de la variables *gre* es prácticamente 1, lo que quiere decir que el que tengas un punto más en el exámen no es significativo para que te admitan en la universidad.

La odd ratio de la variable *gpa* es 2,17, lo que quiere decir que si tienes un punto más de media es $2,17$ veces más probable que te admitan en la universidad.

Por último, la odd ratio de la variable *rank* es 0,57, lo que quiere decir que si la institución tiene un punto más de prestigio, es un 57 % menos probable que te admitan en la universidad.