Aplicaciones del modelo de regresión logística: analisis de relaciones entre variables

Contents

1 Datos

```
d = read.table("datos/titanic.txt", header = T)
str(d)
```

```
## 'data.frame':    2201 obs. of  4 variables:
##  $ Class   : chr  "First" "First" "First" "First" ...
##  $ Age     : chr  "Adult" "Adult" "Adult" "Adult" ...
##  $ Sex     : chr  "Male" "Male" "Male" "Male" ...
##  $ Survived: int  1 1 1 1 1 1 1 1 1 1 ...
```

donde Survived = 1 indica que el pasajero sobrevivió, y Survived = 0 indica que el pasajero no sobrevivió.

Primero convertimos a factor las variables:

```
d$Class = factor(d$Class)
d$Age = factor(d$Age)
d$Sex = factor(d$Sex)
```

2 Análisis

2.1 Probabilidad de supervivencia de la tripulacion frente a los pasajeros

```
m1 = glm(Survived ~ Class, data = d, family = binomial)
summary(m1)
```

```
##
## Call:
## glm(formula = Survived ~ Class, family = binomial, data = d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3999  -0.7623  -0.7401   0.9702   1.6906
##
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.15516    0.07876 -14.667  < 2e-16 ***
## ClassFirst   1.66434    0.13902  11.972  < 2e-16 ***
## ClassSecond  0.80785    0.14375   5.620 1.91e-08 ***
## ClassThird   0.06785    0.11711   0.579    0.562
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2769.5  on 2200  degrees of freedom
## Residual deviance: 2588.6  on 2197  degrees of freedom
## AIC: 2596.6
##
## Number of Fisher Scoring iterations: 4
```

El nivel de referencia de Class es tripulación (Crew). Por tanto, β_1 representa diferencias en la probabilidad de supervivencia entre Crew y First; β_2 representa diferencias entre Crew y Second; β_3 representa diferencias entre Crew y Third.

Los pvalores para β_1 y β_2 son menores que 0.05, por lo que son distintas de cero, luego hay diferencias en la probabilidad de supervivencia. Como $\hat{\beta}_1 = 1.664344 > 0$, la probabilidad de supervivencia de la primera clase es mayor que la de la tripulación; lo mismo para la segunda clase.

El pvalor de β_3 es mayor que 0.05, luego $\beta_3 = 0$, no hay diferencias entre tripulación y tercera clase.

Estos resultados se pueden comprobar numéricamente:

```
(predP = predict(m1, newdata = data.frame(Class = c("Crew", "First", "Second", "Third")), type = "respon
```

```
##         1         2         3         4
## 0.2395480 0.6246154 0.4140351 0.2521246
```

con intervalos de confianza

```
predL = predict(m1, newdata = data.frame(Class = c("Crew", "First", "Second", "Third")), type = "link",
#
alfa = 0.05
Lp = predL$fit - qnorm(1-alfa/2)*predL$se.fit
Up = predL$fit + qnorm(1-alfa/2)*predL$se.fit
#
data.frame(pred = predP, confintL = exp(Lp)/(1+exp(Lp)), confintP = exp(Up)/(1+exp(Up)))
```

```
##        pred  confintL  confintP
## 1 0.2395480 0.2125668 0.2687850
## 2 0.6246154 0.5706887 0.6756185
## 3 0.4140351 0.3582390 0.4721282
## 4 0.2521246 0.2214588 0.2854798
```

Donde podemos comprobar que algunos intervalos se solapan y otros no.

2.2 Probabilidad de supervivencia de primera y segunda clase

```
d$Class = relevel(d$Class, ref = "Second")
m2 = glm(Survived ~ Class, data = d, family = binomial)
summary(m2)
```

```
##
## Call:
## glm(formula = Survived ~ Class, family = binomial, data = d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3999  -0.7623  -0.7401   0.9702   1.6906
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.3473     0.1203  -2.888  0.00388 **
## ClassCrew    -0.8078     0.1438  -5.620 1.91e-08 ***
## ClassFirst    0.8565     0.1661   5.157 2.51e-07 ***
## ClassThird   -0.7400     0.1482  -4.992 5.98e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2769.5  on 2200  degrees of freedom
## Residual deviance: 2588.6  on 2197  degrees of freedom
## AIC: 2596.6
##
## Number of Fisher Scoring iterations: 4
```

El pvalor de β_2 es menor que 0.05, luego hay diferencias entre probabilidades de supervivencia. En concreto, como $\beta_2 > 0$, la probabilidad de supervivencia de la primera clase es mayor.

2.3 Probabilidad de supervivencia de mujeres de tripulación y mujeres pasajeras

Podemos utilizar relevel de nuevo:

```r
d$Class = relevel(d$Class, ref = "Crew")
m3a = glm(Survived ~ Class * Sex, data = d, family = binomial)
summary(m3a)
```

```
##
## Call:
## glm(formula = Survived ~ Class * Sex, family = binomial, data = d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6797  -0.7099  -0.6155   0.5115   1.9842
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.89712    0.61914   3.064  0.00218 **
## ClassSecond          0.07053    0.68630   0.103  0.91815
## ClassFirst           1.66535    0.80026   2.081  0.03743 *
## ClassThird          -2.06075    0.63551  -3.243  0.00118 **
## SexMale             -3.14690    0.62453  -5.039 4.68e-07 ***
## ClassSecond:SexMale -0.63882    0.72402  -0.882  0.37760
## ClassFirst:SexMale  -1.05911    0.81959  -1.292  0.19627
## ClassThird:SexMale   1.74286    0.65139   2.676  0.00746 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2769.5  on 2200  degrees of freedom
## Residual deviance: 2163.7  on 2193  degrees of freedom
## AIC: 2179.7
##
## Number of Fisher Scoring iterations: 6
```

Aunque se puede utilizar relevel, quiero mantener el orden de los niveles Crew-First-Seconf-Third. Para ello:

```r
d$Class = factor(d$Class, levels = c("Crew","First","Second","Third"))
m3 = glm(Survived ~ Class * Sex, data = d, family = binomial)
summary(m3)
```

```
##
## Call:
## glm(formula = Survived ~ Class * Sex, family = binomial, data = d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6797  -0.7099  -0.6155   0.5115   1.9842
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.89712    0.61914   3.064  0.00218 **
## ClassFirst           1.66535    0.80026   2.081  0.03743 *
## ClassSecond          0.07053    0.68630   0.103  0.91815
## ClassThird          -2.06075    0.63551  -3.243  0.00118 **
## SexMale             -3.14690    0.62453  -5.039 4.68e-07 ***
## ClassFirst:SexMale  -1.05911    0.81959  -1.292  0.19627
## ClassSecond:SexMale -0.63882    0.72402  -0.882  0.37760
## ClassThird:SexMale   1.74286    0.65139   2.676  0.00746 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2769.5  on 2200  degrees of freedom
## Residual deviance: 2163.7  on 2193  degrees of freedom
## AIC: 2179.7
##
## Number of Fisher Scoring iterations: 6
```

Como vemos, los resultados son iguales. Se han creado las siguientes variables auxiliares (entre paréntesis la abreviatura)

- ClassFirst (F)
- ClassSecond (S)
- ClassThird (T)
- SexMale (M)

El modelo que se ha estimado es:

$$P(Y_i = 1) = \frac{exp(\beta_0 + \beta_1 F_i + \beta_2 S_i + \beta_3 T_i + \beta_4 M_i + \beta_5 F_i \cdot M_i + \beta_6 S_i \cdot M_i + \beta_7 T_i \cdot M_i)}{1 + exp(\beta_0 + \beta_1 F_i + \beta_2 S_i + \beta_3 T_i + \beta_4 M_i + \beta_5 F_i \cdot M_i + \beta_6 S_i \cdot M_i + \beta_7 T_i \cdot M_i)}$$

Para mujeres de tripulación F = 0, S = 0, T = 0, M = 0. Por tanto el modelo es:

$$P(Y_i = 1) = \frac{exp(\beta_0)}{1 + exp(\beta_0)}$$

Para mujeres de tercera clase, por ejemplo, F = 0, S = 0, T = 1, M = 0. Por tanto el modelo es:

$$P(Y_i = 1) = \frac{exp(\beta_0 + \beta_3)}{1 + exp(\beta_0 + \beta_3)}$$

Por tanto, β_3 modela las diferencias de probabilidad de supervivencia. Como pvalor de $\beta_3 < 0.05$, las probabilidades no son iguales. Como $\beta_3 < 0$, las probabilidades de las mujeres de tripulación son mayores que las mujeres de tercera clase.

2.4 Probabilidad de supervivencia de hombres de tripulación y hombres pasajeros

Utilizamos el mismo modelo del apartado anterior.

Para hombres de tripulación F = 0, S = 0, T = 0, M = 1. Por tanto el modelo es:

$$P(Y_i = 1) = \frac{exp(\beta_0 + \beta_4)}{1 + exp(\beta_0 + \beta_4)}$$

Para hombres de tercera clase, F = 0, S = 0, T = 1, M = 1. Por tanto el modelo es:

$$P(Y_i = 1) = \frac{exp(\beta_0 + \beta_3 + \beta_4 + \beta_7)}{1 + exp(\beta_0 + \beta_3 + \beta_4 + \beta_7)}$$

Por tanto, $\beta_3 + \beta_7$ modela las diferencias de probabilidad de supervivencia. Se formula el siguiente contraste:

$$\beta_3 + \beta_7 = 0 \beta_3 + \beta_7 \neq 0$$

El estadístico del contraste es:

$$\hat{\beta}_3 + \hat{\beta}_7 \sim N(\beta_3 + \beta_7, se(\hat{\beta}_3 + \hat{\beta}_7))$$

donde

$$se(\hat{\beta}_3 + \hat{\beta}_7) = \sqrt{V_{44} + V_{88} + 2V_{48}} V = Var(\hat{\beta})$$

```
source("logit_funciones.R")
H = logit_hess(coef(m3), model.matrix(m3))
V = -solve(H)
(se = sqrt(V[4,4] + V[8,8] + 2*V[4,8]))
```

```
## [1] 0.1429482
```

El estadístico se reescribe como

$$z = \frac{\hat{\beta}_3 + \hat{\beta}_7}{se(\hat{\beta}_3 + \hat{\beta}_7)}$$

```
(z = (coef(m3)[4] + coef(m3)[8])/se)
```

```
## ClassThird
##  -2.223786
```

```
(pvalor = 2*(1-pnorm(abs(z))))
```

```
## ClassThird
## 0.02616282
```

Como es menor que 0.05 SI hay diferencias en la probabilidad de superviviencia. Como $\hat{\beta}_3 + \hat{\beta}_7 = -2.0607494 + 1.7428632 = -0.3178862 < 0$, la probabilidad de supervivencia de los hombres de tripulación es mayor que la de los hombres de tercera clase.