# Regresion Logística

Yanink Neried Caro Vega y Lizbeth Naranjo Albarrán

## Ejercicio 5.35 del Libro de Agresti (2015)

*The New York Times reported results of a study on the effects of AZT in slowing the development of AIDS symptoms (February 15, 1991). Veterans whose immune symptoms were beginning to falter after infection with HIV were randomly assigned to receive AZT immediately or wait until their T cells showed severe immune weakness. During the 3-year study, of those who received AZT, 11 of 63 black subjects and 14 of 107 white subjects developed AIDS symptoms. Of those who did not receive AZT, 12 of 55 black subjects and 32 of 113 white subjects developed AIDS symptoms. Use model building, including checking fit and interpreting effects and inference, to analyze these data.*

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Wiley Series in Probability and Statistics. Wiley.

Pregunta de interés:
**¿Existe el efecto significativo del tratamiento de AZT en el desarrollo de SIDA?**

### Datos

| Tratamiento | Raza | Si SIDA | No SIDA | Total |
|---|---|---|---|---|
| Reciben AZT inmediatamente | Negra | 11 | 52 | 63 |
| | Blanca | 14 | 93 | 107 |
| No reciben AZT, esperar | Negra | 12 | 43 | 55 |
| | Blanca | 32 | 81 | 113 |
| | | 69 | 269 | 338 |

### Modelo con datos NO agrupados (respuesta 0 y 1)

$$Y_i \sim Bernoulli(p_i)$$

- $Y_i$ es el número de casos que Sí desarrollaron síntomas de SIDA.
- Considerando que se tienen $n$ casos expuestos (Total).

```
sida = c(rep(1,11),rep(0,52),rep(1,14),rep(0,93),rep(1,12),rep(0,43),rep(1,32),rep(0,81))
tratamiento = c(rep("Recibir AZT",63+107),rep("Esperar",55+113))
raza = c(rep("Negra",63),rep("Blanca",107),rep("Negra",55),rep("Blanca",113))
table(sida,tratamiento,raza)
```

```
## , , raza = Blanca
##
##     tratamiento
```

```
## sida Esperar Recibir AZT
##    0      81          93
##    1      32          14
##
## , , raza = Negra
##
##     tratamiento
## sida Esperar Recibir AZT
##    0      43          52
##    1      12          11
```

```
head(cbind(sida,tratamiento,raza),3)
```

```
##      sida tratamiento   raza
## [1,] "1"  "Recibir AZT" "Negra"
## [2,] "1"  "Recibir AZT" "Negra"
## [3,] "1"  "Recibir AZT" "Negra"
```

```
tail(cbind(sida,tratamiento,raza),3)
```

```
##        sida tratamiento raza
## [336,] "0"  "Esperar"   "Blanca"
## [337,] "0"  "Esperar"   "Blanca"
## [338,] "0"  "Esperar"   "Blanca"
```

```
length(sida)
```

```
## [1] 338
```

```
sum(sida)
```

```
## [1] 69
```

```
# Modelos univariados
modelo.raza <- glm(sida ~ raza, family=binomial(link="logit"))
summary(modelo.raza)
```

```
##
## Call:
## glm(formula = sida ~ raza, family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6849  -0.6849  -0.6849  -0.6585   1.8084
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.33041    0.16579  -8.025 1.02e-15 ***
## razaNegra   -0.08797    0.28547  -0.308    0.758
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 342.12  on 337  degrees of freedom
## Residual deviance: 342.02  on 336  degrees of freedom
## AIC: 346.02
##
## Number of Fisher Scoring iterations: 4
```

```
modelo.trat <- glm(sida ~ tratamiento, family=binomial(link="logit"))
summary(modelo.trat)
```

```
##
## Call:
## glm(formula = sida ~ tratamiento, family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7793  -0.7793  -0.5640  -0.5640   1.9580
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.0361     0.1755  -5.904 3.54e-09 ***
## tratamientoRecibir AZT   -0.7218     0.2787  -2.590  0.00961 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 342.12  on 337  degrees of freedom
## Residual deviance: 335.19  on 336  degrees of freedom
## AIC: 339.19
##
## Number of Fisher Scoring iterations: 4
```

```
# Modelo multivariado
modelo2 <- glm(sida ~ tratamiento+raza, family=binomial(link="logit"))
summary(modelo2)
```

```
##
## Call:
## glm(formula = sida ~ tratamiento + raza, family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7854  -0.7668  -0.5694  -0.5549   1.9733
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -1.01809    0.19851  -5.129 2.92e-07 ***
## tratamientoRecibir AZT -0.71946    0.27898  -2.579  0.00991 **
## razaNegra               -0.05548    0.28861  -0.192  0.84755
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 342.12  on 337  degrees of freedom
## Residual deviance: 335.15  on 335  degrees of freedom
## AIC: 341.15
##
## Number of Fisher Scoring iterations: 4
```

**Interpretación**

Momios del Tratamiento AZT:

$$\begin{aligned} \frac{p_{azt}}{1-p_{azt}} &= e^{\beta_0+\beta_1} \\ \frac{\mathbb{P}[\text{desarrollar SIDA}|RecibirAZT]}{\mathbb{P}[\text{No desarrollar SIDA}|RecibirAZT]} &= e^{\beta_0+\beta_1} \end{aligned}$$

```
exp(-1.01809-0.71946)   ### 0.175951   factor de proteccion (no de riesgo)
```

```
## [1] 0.175951
```

indicando que entre los que Sí recibieron AZT inmediatamente, existen 17 que desarrollaron SIDA por cada 100 que no desarrollaron la enfermedad. Todas las demás variables en su categoría de referencia.

Momios para los que No reciben AZT de inmediato (categoria de referencia), y esperan hasta que sus células T muestren una debilidad inmunológica grave:

$$\begin{aligned} \frac{p_{esperar}}{1-p_{esperar}} &= e^{\beta_0} \\ \frac{\mathbb{P}[\text{desarrollar SIDA}|Esperar]}{\mathbb{P}[\text{No desarrollar SIDA}|Esperar]} &= e^{\beta_0} \end{aligned}$$

```
exp(-1.01809)   ### 0.3612843   factor de proteccion (no de riesgo)
```

```
## [1] 0.3612843
```

indicando que entre los que No recibieron tratamiento inmediatamente, existen 36 que desarrollaron SIDA por cada 100 que no desarrollaron la enfermedad. Todas las demás variables en su categoría de referencia.

Cociente de momios (*odds ratio*):

$$\frac{\frac{p_{azt}}{1-p_{azt}}}{\frac{p_{esperar}}{1-p_{esperar}}} = e^{\beta_1}$$

```r
exp(-0.71946)   ### 0.4870152  ###  = 0.175951/0.3612843
```

```
## [1] 0.4870152
```

indicando que ante el tratamiento de AZT inmediato, la posibilidad de desarrollar SIDA es 0.48 veces más que en el caso de esperar a que sus células T muestren una debilidad inmunológica grave. Todas las demás variables en su categoría de referencia.

Equivale a que los odds de desarrollar SIDA ante el tratamiento de AZT inmediato es 0.48 veces mayor (52% menor) que los odds de desarrollar SIDA que en el caso de esperar a que sus células T muestren una debilidad inmunológica grave.

Cociente de momios (*odds ratio*):

$$\frac{\frac{p_{esperar}}{1-p_{esperar}}}{\frac{p_{azt}}{1-p_{azt}}} = e^{-\beta_1}$$

```r
exp(0.71946)   ### 2.053324  ###  = 0.3612843/0.175951
```

```
## [1] 2.053324
```

Equivale a que los odds de desarrollar SIDA en el caso de esperar a que sus células T muestren una debilidad inmunológica grave es 2.05 veces mayor (105% mayor) que los odds de desarrollar SIDA ante el tratamiento de AZT inmediato.

### Sensibilidad, Especificidad y Curva ROC

```r
library(AUC)
```

```
## AUC 0.3.0

## Type AUCNews() to see the change log and ?AUC to get an overview.
```

```r
pred2 <- predict(modelo2,type="response")   ### Predecir P[Y=1]
Ypred2 <- ifelse(pred2>0.15,1,0)   ### Predecir Y, considerando como punto de corte 0.5
table(sida,Ypred2)   ### Tabla de observaciones vs. predicciones
```

```
##     Ypred2
## sida   0   1
##    0 145 124
##    1  25  44
```

```r
VP <- sum(sida==1 & Ypred2==1)
VN <- sum(sida==0 & Ypred2==0)
FP <- sum(sida==0 & Ypred2==1)
FN <- sum(sida==1 & Ypred2==0)
VP/(VP+FN)   ### sensibilidad
```

```
## [1] 0.6376812
```

```
VN/(VN+FP)   ### especificidad
```

```
## [1] 0.5390335
```

```
sens2 <- sensitivity(pred2,factor(sida))   ### Calcula la sensibilidad
head(sens2$cutoffs)   ### puntos de corte para calcular la sensibilidad
```

```
##                   338        225        170         63
## 1.00000000 0.66863905 0.50591716 0.18934911 0.00295858 0.00000000
```

```
head(sens2$measure)   ### valores de la sensibilidad
```

```
## [1] 0.0000000 0.4637681 0.6376812 0.8405797 1.0000000 1.0000000
```

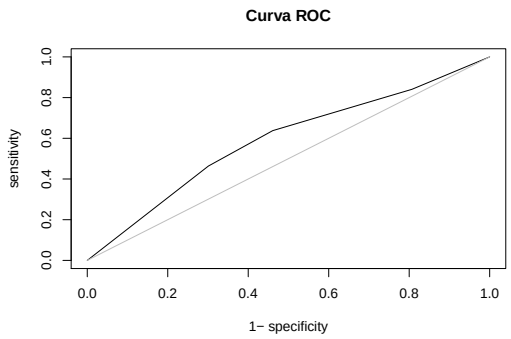
```
spec2 <- specificity(pred2,factor(sida))   ### Calcula la especificidad
head(spec2$cutoffs)   ### puntos de corte para calcular la especificidad
```

```
##                  338       225       170        63
## 1.0000000 1.0000000 0.6656805 0.5029586 0.1863905 0.0000000
```

```
head(spec2$measure)   ### valores de la especificidad
```

```
## [1] 1.0000000 0.6988848 0.5390335 0.1933086 0.0000000 0.0000000
```

```
plot(roc(pred2,factor(sida)),main="Curva ROC")
```




```
auc(roc(pred2,factor(sida)))   ### Area bajo la curva ROC
```

```
## [1] 0.5912936
```

**Modelo con datos agrupados**

Esto se puede representar como:

$$Y_i \sim Binomial(n_i, p_i)$$

- $Y_i$ es el número de casos que Sí desarrollaron síntomas de SIDA.
- Considerando que se tienen $n_i$ casos expuestos (Total).
- Para el conjunto de características $x_i$ (recibe AZT o no, raza negra o blanca).
- Cuya probabilidad de presentar síntomas de SIDA es $p_i$.

El modelo de regresión logística implica que la forma en cómo se relaciona $p_i$ con $x_i$ es:

$$p_i = \frac{e^{x_i\beta}}{1+e^{x_i\beta}} \qquad o \qquad \log\left(\frac{p_i}{1-p_i}\right) = x_i\beta$$

```
sidaG = c(11,14,12,32)
totalG = c(63,107,55,113)
tratamientoG = factor(c("Recibir AZT","Recibir AZT","Esperar","Esperar"))
razaG = factor(c("Negra","Blanca","Negra","Blanca"))
modeloG <- glm(cbind(sidaG,totalG-sidaG) ~ tratamientoG+razaG, family=binomial(link="logit"))
summary(modeloG)
```

```
##
## Call:
## glm(formula = cbind(sidaG, totalG - sidaG) ~ tratamientoG + razaG,
##     family = binomial(link = "logit"))
##
## Deviance Residuals:
##       1        2        3        4
##  0.7035  -0.5547  -0.6326   0.4253
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.01809    0.19851  -5.129 2.92e-07 ***
## tratamientoGRecibir AZT -0.71946    0.27898  -2.579  0.00991 **
## razaGNegra               -0.05548    0.28861  -0.192  0.84755
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8.3499  on 3  degrees of freedom
## Residual deviance: 1.3835  on 1  degrees of freedom
## AIC: 24.86
##
## Number of Fisher Scoring iterations: 4
```