Biostatistics Exercises

www.nataliejulian.com

Scoring

Las bases de datos `score_train.txt` y `score_test.txt` contienen información sobre la adherencia que tiene un grupo de pacientes respecto a cierto tratamiento (es decir, qué tan responsables son para seguir un tratamiento médico, especialmente cuando se les asignan medicamentos). Una compañía farmacéutica está interesada en entregar un programa completo de medicamentos a pacientes que adhieren de buena manera al tratamiento. Suponga que existen 5 covariables que definen el perfil de adherencia y la variable respuesta es adherencia (1 si tiene un historial positivo de adherencia al tratamiento, 0 si no).

Suponga que la compañía farmacéutica le encarga a usted evaluar a individuos de la base de datos `score_test.txt`. Usted debe asignar un puntaje de adherencia a cada individuo y decidir si se le otorgará el programa de medicamentos, tal que si el puntaje de adherencia del individuo es menor al puntaje de corte, entonces no se otorga el programa, y, en caso contrario, sí se otorga. Para decidir, realice lo siguiente:

a) Ajuste una regresión logística con los datos de entrenamiento. Considere adherencia como variable respuesta y el resto como variables explicativas.

```
spec_tbl_df [700 x 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ "adherencia": logi [1:700] TRUE FALSE TRUE FALSE TRUE TRUE ...
 $ "x1"        : num [1:700] -1.018 -0.876 -0.743 0.605 0.647 ...
 $ "x2"        : num [1:700] 0.323 -2.456 1.907 0.474 1.546 ...
 $ "x3"        : num [1:700] -1.2111 0.242 1.8238 -0.4887 -0.0605 ...
 $ "x4"        : num [1:700] 0.0499 -0.2173 -0.5024 0.5246 0.3921 ...
 $ "x5"        : num [1:700] 0.241 -1.379 0.26 -2.405 -0.452 ...
 - attr(*, "spec")=
  .. cols(
  ..   `"adherencia"` = col_logical(),
  ..   `"x1"` = col_double(),
  ..   `"x2"` = col_double(),
  ..   `"x3"` = col_double(),
  ..   `"x4"` = col_double(),
```

```
  ..   `"x5"` = col_double()
  .. )

> score_train$adhind<-factor(ifelse(score_train$adherencia==TRUE, "1","0"))
> score<-glm(adhind ~ ., family = binomial,data = score_train[,-1])
> summary(score)  #Todas las variables son significativas

Call:
glm(formula = adhind ~ ., family = binomial, data = score_train[,
    -1])

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.11009  -0.19611   0.01005   0.21813   2.49037

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.0474     0.1710   6.126 9.01e-10 ***
x1            0.9775     0.1736   5.630 1.81e-08 ***
x2            3.3107     0.3102  10.673  < 2e-16 ***
x3            2.3575     0.2373   9.933  < 2e-16 ***
x4            2.7470     0.2698  10.180  < 2e-16 ***
x5            2.1485     0.2298   9.349  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 957.20  on 699  degrees of freedom
Residual deviance: 299.99  on 694  degrees of freedom
AIC: 311.99

Number of Fisher Scoring iterations: 7
```

b) Con los datos de testeo encuentre el puntaje de corte bajo el criterio *Distance to corner*. Este criterio está basado en la distancia a la esquina superior izquierda de la curva ROC y está definido para cada punto de corte. La Distance to the corner está definida como

$$d = \sqrt{(1 - sensitivity)^2 + (1 - specificity)^2}$$

El punto de corte que minimice esta distancia será considerado como puntaje de corte para otorgar el programa de medicamentos. Explique, con sus palabras, por

qué este criterio puede ser considerado como un punto de corte para el puntaje de adherencia.

Esta forma de calcular el punto de corte puede ser utilizado porque la minimización de la funcion *Distance to corner* pone en juego tanto la sensibilidad como la especificidad y además la sensibilidad y especificidad se encuentran en el intervalo [0,1] al ser tasas, por lo que el punto de corte caerá dentro del intervalo [0,1]:

```
> distancecorner <- function (x, y) {
+    sqrt(((1-x)^2 + (1-y)^2)) }
> x <- seq(0, 1, length= 30)
> y <- x
> z <- outer(x, y, distancecorner)
> persp(x, y, z, main="Distance to corner y punto óptimo",
+ zlab = "Distance ro Corner", xlab="Especificidad",
+ ylab="Sensibilidad",
+ theta = -20, phi = 0, col = "orchid", shade = 0.5, d=0.1)
```

Cálculo del punto de corte:

```
> score_test$adhind<-factor(ifelse(score_test$adherencia==TRUE, "1","0"))
> Matrix<-cbind(score_test[, 2:6])
> pred<-exp(coef(score)[1]+coef(score)[2]*Matrix[,2]+
+             coef(score)[3]*Matrix[,3]+coef(score)[4]*Matrix[,4]+
+             coef(score)[5]*Matrix[,5])/(1+exp(coef(score)[1]+
+                                                 coef(score)[2]*Matrix[,2]+
+             coef(score)[3]*Matrix[,3]+coef(score)[4]*Matrix[,4]+
+             coef(score)[5]*Matrix[,5]))
> #Tomaré una secuencia del punto de corte
>
> secuencia<-seq(0,1, by=0.005)
> Sens<-rep(0, length(secuencia))
> Esp<-rep(0, length(secuencia))
> for(i in 1:length(secuencia)){
+   values<-factor(ifelse(pred<secuencia[i], "0","1"))
+
+   datos<-as.vector(table(values,score_test$adhind))
+   Sens[i]<-datos[4]/(datos[4]+datos[3])
+   Esp[i]<-datos[1]/(datos[1]+datos[2])}
> d<-cbind(secuencia, Sens, Esp, sqrt((1-Sens)^{2}+(1-Esp)^{2}))
> colnames(d)<-c("d","Sensibilidad","Especificidad", "Corner")
> head(d[order(d[,4]),],10)

          d Sensibilidad Especificidad    Corner
 [1,] 0.560    0.8148148     0.7971014 0.2747023
```

```
 [2,] 0.565    0.8148148     0.7971014 0.2747023
 [3,] 0.570    0.8148148     0.7971014 0.2747023
 [4,] 0.575    0.8148148     0.7971014 0.2747023
 [5,] 0.580    0.8148148     0.7971014 0.2747023
 [6,] 0.585    0.8148148     0.7971014 0.2747023
 [7,] 0.590    0.8148148     0.7971014 0.2747023
 [8,] 0.535    0.8209877     0.7898551 0.2760549
 [9,] 0.540    0.8209877     0.7898551 0.2760549
[10,] 0.545    0.8209877     0.7898551 0.2760549

> d[which(d[,4]==min(d[,4], na.rm=TRUE)),] #Se pudiera elegir cualquiera

         d Sensibilidad Especificidad    Corner
[1,] 0.560    0.8148148     0.7971014 0.2747023
[2,] 0.565    0.8148148     0.7971014 0.2747023
[3,] 0.570    0.8148148     0.7971014 0.2747023
[4,] 0.575    0.8148148     0.7971014 0.2747023
[5,] 0.580    0.8148148     0.7971014 0.2747023
[6,] 0.585    0.8148148     0.7971014 0.2747023
[7,] 0.590    0.8148148     0.7971014 0.2747023

> (puntodecorte<-unname(d[which.min(d[,4]),1]))

[1] 0.56

>
```

c) Con el punto de corte encontrado en 2., decida a qué individuos les otorgaría el programa y a cuáles no. Para mostrar los individuos a los cuáles les otorgaría el programa, defina una nueva columna en la base de datos `score_test.txt` tal que 1=Se otorgará programa y 0=No se otorgará programa (La base de datos con dicha columna creada es parte de la entrega de la tarea.)

```
> score_test$Decision<-factor(ifelse(pred<puntodecorte, "0","1"))
> # 0: No se entrega programa si prob<punto de corte
>
> # 1: Se otorgara programa si prob>= punto de corte
>
> head(score_test)

# A tibble: 6 x 8
  adherencia       x1     x2     x3     x4     x5 adhind Decision
  <lgl>         <dbl>  <dbl>  <dbl>  <dbl>  <dbl> <fct>  <fct>
1 TRUE        1.12     0.290 -0.441 -1.10   0.985 1      0
2 TRUE       -0.00552 -0.365  0.892  0.837 -1.24  1      1
```

```
3 TRUE       -0.0945   0.358  0.435  1.67   0.881 1      1
4 FALSE      -0.134   -1.38  -1.16   0.714 -0.885 0      0
5 TRUE       -0.0660   1.53  -0.506 -0.743  1.08  1      1
6 TRUE        1.25     0.238  1.61   0.775 -0.123 1      1

> table(score_test$Decision)

  0   1
140 160

> #De los 300 pacientes a 160 se le otorga el programa.
>
>
> write.csv(score_test, "score_test2.csv", row.names = FALSE)
> #Base de datos con decision
```