Actividad 3: Modelización predictiva

Solución

21 de diciembre 2018

Índice

En esta actividad se usará el fichero **Fumadores_clean_5Y_1.csv** ya preparado, es decir, después del preproceso que se ha realizado en la primera actividad.

Recuerde que este archivo almacena los datos de una investigación médica sobre la capacidad pulmonar de varias personas, con el objetivo de estudiar si los hábitos de salud y los hábitos como fumadores influencian la capacidad pulmonar. Para realizar el estudio se recogió una muestra de 300 personas. A cada persona, se le preguntó a través de un cuestionario su género, hábitos de deporte, si era fumadora, y en caso de que lo fuera, cuántos cigarrillos al día de promedio fumaba y los años que hacía que fumaba. Además, se midió la capacidad pulmonar de cada persona a partir de un test de aire expulsado, desde donde se tomó como capacidad pulmonar la medida FEF (forced expiratory flow), que es la velocidad del aire saliendo del pulmón durante la porción central de una espiración forzada. Se mide en litros / segundo. Además, se incluye la variable `PC5Y` que es la capacidad pulmonar de cada persona medida al cabo de 5 años de realizar el primer test. Se asume que la persona no ha cambiado sus condiciones personales significativamente en este tiempo.

Otros datos personales recogidos son: la altura, peso y ciudad donde vive.

Esta base de datos contiene 300 registros y 10 variables. Las variables son Sex, Sport, Years, Cig, PC, City, Weight, Age, Height, PC5Y.

---

1. Modelo de regresión lineal

Primeramente, estudiaremos la posible asociación entre la capacidad pulmonar y algunas características de cada individuo.

1.1. Modelo de regresión lineal múltiple (regresores cuantitativos)

Estimar por mínimos cuadrados ordinarios un modelo lineal que explique la capacidad pulmonar (PC) de un individuo en función de tres factores cuantitativos: el peso (Weight), el número de cigarrillos que fuma al dia

(Cig), y el número de años que hace que fuma (Years).

Evaluar la bondad de ajuste a través del coeficiente de determinación (R^2). Podéis usar la instrucción de R `lm`.

Ademas, evaluar si alguno de los regresores tiene influencia significativa (p-valor del contraste individual inferior al 5 %).

```
#Estimacion del modelo

Model.1.1<- lm(PC~Years+Cig+Weight, data=mydata )
summary(Model.1.1)
```

```
##
## Call:
## lm(formula = PC ~ Years + Cig + Weight, data = mydata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97891 -0.18424 -0.01939  0.19799  0.78591
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.677888   0.284701  12.918   <2e-16 ***
## Years       -0.023139   0.001583 -14.613   <2e-16 ***
## Cig         -0.032711   0.001923 -17.008   <2e-16 ***
## Weight       0.001283   0.004178   0.307    0.759
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2734 on 296 degrees of freedom
## Multiple R-squared:  0.812,  Adjusted R-squared:  0.8101
## F-statistic: 426.3 on 3 and 296 DF,  p-value: < 2.2e-16
```

El coeficiente de la bondad de ajuste es 0.812 y el coeficiente ajustado es: 0.8101. Además, se observa que el test global de la regresión es significativo.

Por otra parte, han sido significativos los test parciales sobre los coeficientes de los regresores Years, Cig.

Observa que, a diferencia de Weight, no se ha añadido al modelo de regresión la variable Height, ¿desde el punto de vista de la calidad del modelo de regresión, puedes indicar una razón que justifique el no hacerlo?

Dado que la multicolinealidad entre las variables explicativas es un factor de inestabilidad en la estimación de los coeficientes de regresión. Es interesante explorar la matriz de correlación entre regresores.

```
is_number <- sapply(mydata,is.numeric)
a <-cor(mydata[,is_number])
a
```

```
##              Years        Cig          PC      Weight         Age
## Years   1.00000000  0.6018955 -0.78933308 -0.01794329  0.31674438
## Cig     0.60189555  1.0000000 -0.82221450 -0.13075821 -0.17408261
## PC     -0.78933308 -0.8222145  1.00000000  0.08710433 -0.04524602
## Weight -0.01794329 -0.1307582  0.08710433  1.00000000  0.16685580
## Age     0.31674438 -0.1740826 -0.04524602  0.16685580  1.00000000
## Height -0.04492902 -0.1462693  0.13135340  0.94144547  0.16165046
## PC5Y   -0.79341380 -0.8305974  0.99400386  0.09757945 -0.03032891
##             Height        PC5Y
## Years  -0.04492902 -0.79341380
```

```
## Cig    -0.14626926 -0.83059736
## PC      0.13135340  0.99400386
## Weight  0.94144547  0.09757945
## Age     0.16165046 -0.03032891
## Height  1.00000000  0.14119513
## PC5Y    0.14119513  1.00000000
```

Observa que la correlación entre Weight y Height es un valor muy alto, 0.941. Lo cual ha sugerido no introducir una de las variables en el modelo.

1.2. Modelo de regresión lineal múltiple (regresores cuantitativos y cualitativos)

Estimar por mínimos cuadrados ordinarios un modelo lineal que explique la capacidad pulmonar (PC) de un individuo en función de cinco regresores. Además de los tres anteriores (Years, Cig y Weight) ahora se añade las variables Sex y Sport. Usar como categoría de referencia de la variable Sex la categoria "F" y de la variable Sport la categoria "N" (para ello usar la función `relevel()`). Se pueden definir nuevas variables, SexR y SportR, para esta nueva reordenación.

Evaluar la bondad del ajuste a través del coeficiente de determinación (R^2) y comparar el resultado de este modelo con el obtenido en el apartado 1.1. Podéis usar la instrucción de R `lm` y usar el coeficiente R-cuadrado ajustado en la comparación. Interpretar también el significado de los coeficientes obtenidos y su significación estadística.

```
mydata$SexR=relevel(mydata$Sex, ref = 'F')
mydata$SportR=relevel(mydata$Sport, ref = 'N')


Model.1.2<- lm(PC~Years+Cig+Weight+SportR+SexR, data=mydata )
summary(Model.1.2)
```

```
##
## Call:
## lm(formula = PC ~ Years + Cig + Weight + SportR + SexR, data = mydata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73530 -0.10739 -0.00663  0.10509  0.48296
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.561531   0.236800  15.040  < 2e-16 ***
## Years       -0.022633   0.001065 -21.249  < 2e-16 ***
## Cig         -0.034035   0.001293 -26.315  < 2e-16 ***
## Weight      -0.000601   0.003599  -0.167 0.867511
## SportRE      0.565597   0.032939  17.171  < 2e-16 ***
## SportRR      0.370267   0.031566  11.730  < 2e-16 ***
## SportRS      0.199592   0.025917   7.701 2.11e-13 ***
## SexRM        0.102336   0.027373   3.739 0.000223 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1827 on 292 degrees of freedom
## Multiple R-squared:  0.9172, Adjusted R-squared:  0.9152
## F-statistic: 462.2 on 7 and 292 DF,  p-value: < 2.2e-16
```

```
sel <- which(summary(Model.1.2)$coefficients[-1,4] < 0.05)
sel <- sel + 1
```

El coeficiente de la bondad de ajuste del primer modelo es 0.8101 y del segundo es 0.9152. Por tanto el mejor modelo es el que tiene un coeficiente ajustado superior. Dado que el segundo modelo es mejor podemos concluir que las variables SexR y SportR introducen ciertas diferencias en el modelo predictivo.

Por otra parte, han sido significativos los test parciales sobre los coeficientes de los regresores Years, Cig, SportRE, SportRR, SportRS, SexRM. Siendo las estimaciones de sus coeficientes -0.023, -0.034, 0.566, 0.37, 0.2, 0.102. El signo negativo en los coeficientes indican que dichos coeficientes tienen un efecto de disminución de la variable capacidad pulmonar (PC), en cambio los coeficientes con signo positivo indican un efecto incrementador de la variable PC.

1.3. Efectuar una predicción de la capacidad pulmonar con los dos modelos

Suponer un hombre de Lleida de 30 años de edad que hace deporte regularmente, de peso 68 kg y de altura 175 cm que fuma desde hace 15 años de 10 cigarros al dia.

Realizar la predicción de la capacidad pulmonar (PC) con los dos modelos. Interpretar los resultados.

```
newdata=data.frame(SexR="M", SportR= "R" , Years=15, Cig=10, Weight=68)

p.mod1.1 <- predict(Model.1.1, newdata, interval= c("confidence"))
p.mod1.2 <-predict(Model.1.2, newdata, interval=c("confidence"))
```

La capacidad pulmonar según el primer modelo seria 3.0909 siendo el intervalo de confianza del 95 %: (3.0557, 3.1261).

Para el segundo modelo, la capacidad pulmonar seria 3.3134 siendo el intervalo de confianza del 95 %: (3.2562, 3.3706).

---

2. Modelo de regresión logística

Se desea evaluar la calidad predictiva de la capacidad pulmonar así como de otras variables presentes en el estudio respecto a la predicción de ser fumador. Por tanto, se evaluará la probabilidad de que un individuo sea fumador.

Para evaluar esta probabilidad se aplicará un modelo de regresión logística, donde la variable depediente será una variable binaria que indicará si el individuo es fumador. Se usará la muestra disponible para estimar el modelo con las mismas variables que en el modelo 1.1.

2.1. Estimación de un modelo de regresión logística

El primer paso será crear una variable binaria (smoker) que indique la condición de fumador (smoker = 1) o no fumador (smoker = 0). Estimar el modelo de regresión logística donde la variable dependiente es “smoker” y las explicativas son la capacidad pulmonar (PC), Weight y SexR.

Evaluar si alguno de los regresores tiene influencia significativa (p-valor del contraste individual inferior al 5 %).

```r
mydata$smoker=(mydata$Cig>0)*1

Model.2.1=glm(smoker~PC + Weight+ SexR, family=binomial, data=mydata)
summary(Model.2.1)
```

```
##
## Call:
## glm(formula = smoker ~ PC + Weight + SexR, family = binomial,
##     data = mydata)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.47787  -0.31210  -0.04843   0.04271   3.13797
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 38.51385    7.64106   5.040 4.65e-07 ***
## PC          -9.50804    1.36436  -6.969 3.20e-12 ***
## Weight      -0.09876    0.08155  -1.211    0.226
## SexRM        0.76825    0.64182   1.197    0.231
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 411.06  on 299  degrees of freedom
## Residual deviance: 107.87  on 296  degrees of freedom
## AIC: 115.87
##
## Number of Fisher Scoring iterations: 8
```

```r
sel <- which(summary(Model.2.1)$coefficients[-1,4] < 0.05)
sel <- sel + 1
```

Ha sido significativo el test parcial sobre el coeficiente de PC. Siendo la estimacion de su coeficiente -9.508.

Evaluando los resultados, ¿se puede decir que un individuo con capacidad pulmonar reducida tiene mayor probabilidad de ser fumador?

¿Se puede decir que ser mujer aumenta la probabilidad de ser fumador?

Un individuo con capacidad pulmonar reducida tiene mayor probabilidad de ser fumador ya que el signo negativo de la PC es un factor de "protección" ante el riesgo de fumar. Entonces cuando menos capacidad pulmonar menos protección.

Observar que la variable "SexR" no es significativa, por tanto el modelo no nos dira nada sobre la probabilidad real.

2.2. Predicción en el modelo lineal generalizado (modelo de regresión logística)

Usando el modelo anterior, calculad la probabilidad de ser fumador para un hombre que tiene una capacidad pulmonar de 3.75 l/s, un peso de 68 kg y altura de 175 cm.

```r
newdata=data.frame(SexR="M", PC=3.75, Weight=68)

predict(Model.2.1, newdata, type= "response")
```

```
##          1
## 0.04356199
```

La predicción de la probabilidad de ser fumador para este caso es 0.0436.

2.3. Mejora del modelo

Buscar un modelo mejor al anterior añadiendo más variables explicativas. Se realizarán las siguientes pruebas:

Modelo regresor que añade al anterior la variable edad (Age).

Modelo regresor que añade la variable SportR.

Modelo regresor que añade Age y SportR.

Decidir si se prefiere el modelo inicial o bien uno de los modelos con Age, con SportR, o con ambas. El criterio para decidir el mejor modelo es AIC. Cuanto más pequeño es AIC mejor es el modelo.

Nota: Si al realizar la regresión logística se obtiene un mensaje similar a:

`glm.fit: fitted probabilities numerically 0 or 1 occurred`

este mensaje nos advierte de una convergencia lenta en el proceso iterativo para hallar las estimaciones. No debemos tenerlo en cuenta.

```
Model.2.2a=glm(smoker~PC + Weight + SexR +Age, family=binomial, data=mydata)
summary(Model.2.2a)
```

```
##
## Call:
## glm(formula = smoker ~ PC + Weight + SexR + Age, family = binomial,
##     data = mydata)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.05290  -0.06361  -0.00431   0.01063   2.84865
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  68.66375   14.34683   4.786 1.70e-06 ***
## PC          -14.39062    2.65981  -5.410 6.29e-08 ***
## Weight       -0.11440    0.10560  -1.083    0.279
## SexRM         1.24369    0.89143   1.395    0.163
## Age          -0.29822    0.06635  -4.494 6.98e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 411.062  on 299  degrees of freedom
## Residual deviance:  56.679  on 295  degrees of freedom
## AIC: 66.679
##
## Number of Fisher Scoring iterations: 9
```

```
Model.2.2b=glm(smoker~ PC + Weight +  SexR + SportR, family=binomial, data=mydata)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
summary(Model.2.2b)
```

```
##
## Call:
## glm(formula = smoker ~ PC + Weight + SexR + SportR, family = binomial,
##     data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0920  -0.1017  -0.0111   0.0012   3.4450
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  67.3890    16.7310   4.028 5.63e-05 ***
## PC          -18.5231     3.5122  -5.274 1.34e-07 ***
## Weight       -0.1098     0.1588  -0.691   0.4893
## SexRM         1.7538     1.1346   1.546   0.1221
## SportRE       9.9092     2.1297   4.653 3.27e-06 ***
## SportRR       7.2364     1.7681   4.093 4.26e-05 ***
## SportRS       2.9985     1.2650   2.370   0.0178 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 411.062  on 299  degrees of freedom
## Residual deviance:  40.261  on 293  degrees of freedom
## AIC: 54.261
##
## Number of Fisher Scoring iterations: 9
```

```r
Model.2.2c=glm(smoker ~ PC +  Weight +  SexR + Age +  SportR , family=binomial, data=mydata)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
summary(Model.2.2c)
```

```
##
## Call:
## glm(formula = smoker ~ PC + Weight + SexR + Age + SportR, family = binomial,
##     data = mydata)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.04847  -0.03630  -0.00360   0.00222   2.36582
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  86.9566    25.9884   3.346 0.000820 ***
## PC          -18.7665     4.4594  -4.208 2.57e-05 ***
## Weight       -0.2067     0.2046  -1.010 0.312354
## SexRM         2.5902     1.4862   1.743 0.081361 .
## Age          -0.2912     0.1069  -2.724 0.006451 **
## SportRE       7.2214     1.9630   3.679 0.000234 ***
## SportRR       5.1093     2.1758   2.348 0.018862 *
```

```
## SportRS       2.3166     1.4635   1.583 0.113436
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 411.062  on 299  degrees of freedom
## Residual deviance:  25.559  on 292  degrees of freedom
## AIC: 41.559
##
## Number of Fisher Scoring iterations: 10
```

En el primer modelo AIC es 115.8748. En los nuevos modelos AIC es: 66.679, 54.261 y 41.5587 respectivamente.

Por tanto, nos inclinamos por el Modelo 2.2c. La elección de este modelo como el mejor modelo sugiere que hay diferencias significativas entre las categorias de Sport, y la significación de PC y Age.

2.4. Calidad del ajuste

Calcular la matriz de confusión del mejor modelo del apartado 2.3 suponiendo un umbral de discriminación del 70 %. Observad cuantos falsos negativos hay e interpretar qué es un falso negativo en este contexto. Hacer lo mismo con los falsos positivos.

```
mydata$prob_smoker= predict(Model.2.2c, mydata, type="response")
mydata$pred_smoker <- ifelse(mydata$prob_smoker > 0.7,1,0)
table(mydata$smoker, mydata$pred_smoker)
```

```
##
##       0   1
##   0 167   2
##   1   4 127
```

Hay 4 falsos negativos. Corresponden a individuos fumadores, pero el modelo ha predicho que su probabilidad de ser fumador es inferior a 0.7.

Hay 2 falsos positivos. Corresponden a individuos no fumadores, pero el modelo ha predicho que su probabilidad de ser fumador es superior a 0.7.

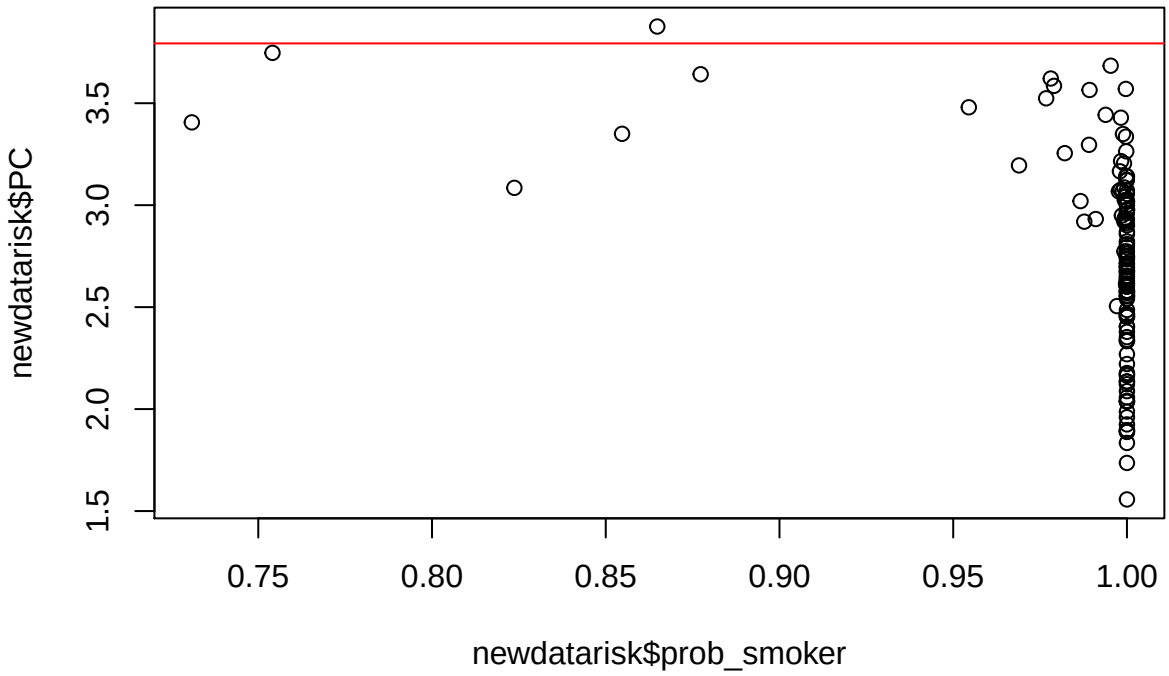
2.5. La selección de los individuos fumadores

Establecer un nivel de probabilidad (umbral de discriminación a partir del cual pensáis que el individuo tiene muchas posibilidades de ser un fumador, por ejemplo podéis escoger el 70 %). Comparar el nivel de probabilidad que da el modelo con el valor de capacidad pulmonar (PC) del individuo. Identificar los individuos que no se comportan según lo esperado, es decir tienen elevada capacidad pulmonar y el modelo los clasifica como fumadores y reportar los valores de probabilidad de ser fumador y de PC. Utilizar como umbral para declarar un individuo con PC elevado el cuartil tercero de la variable PC.

Podéis realizar este estudio gráficamente.

```
mydata$prob_smoker=predict(Model.2.2c, mydata, type="response")
newdatarisk=subset(mydata, prob_smoker>0.7)
Q3 <-quantile(mydata$PC)[4]
PC.anom <- which(newdatarisk$PC>Q3)
```

```r
plot(newdatarisk$prob_smoker, newdatarisk$PC)
abline(h=Q3, col="red")
```



Aparece 1 individuo que tiene PC por encima de 3.79325 y el score basado en el modelo predictivo indica que se trataria de individuos con elevado riesgo de ser fumadores, lo cual seria un comportamiento no esperado pues hemos visto la significación de PC en el modelo predictivo. Deducimos que es la combinación con las otras variables en el modelo la que determina la probabilidad elevada, por encima de 0.7.

2.6. Curva ROC

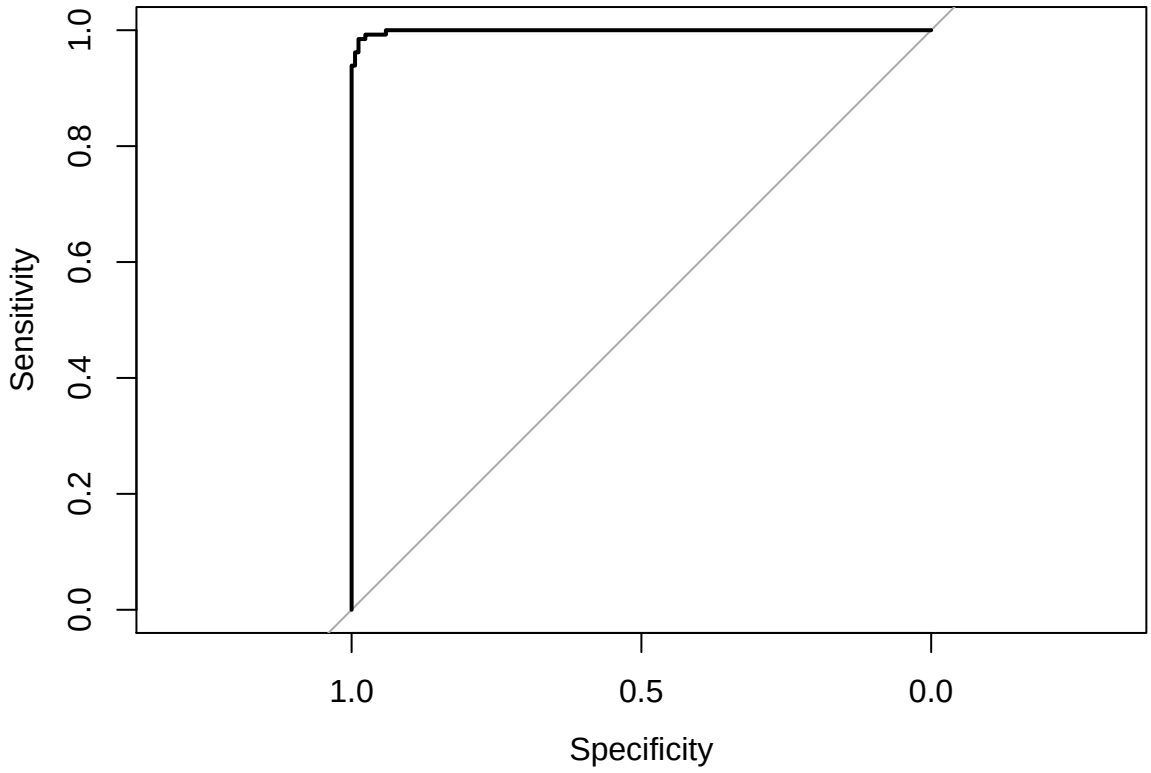
Realizar el dibujo de la curva ROC para representar la calidad del modelo predictivo obtenido. Se puede usar la librería `pROC` y la instrucción `rocy`, finalmente, el plot del objeto resultante. Calcular AUROC usando también este paquete con la función `auc()` donde debéis pasar el nombre del objeto roc.

Interpretar el resultado.

```r
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
##
## Attaching package: 'pROC'
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
g=roc(mydata$smoker,mydata$prob_smoker, data=mydata)
plot(g)
```




```
auc(g)
```

```
## Area under the curve: 0.999
```

AUROC es 0.999.

El modelo logístico tiene un gran poder predictivo, ya que tiene un AUROC elevado, 0.999.

No obstante, seria aconsejable evaluar el modelo logístico bajo la perspectiva de la validación cruzada para evitar un posible problema de sobreajuste a los datos.