2025

Predicción de miopía con R versión ampliada

Fundamentos para la Ciencia de Datos

Miranda, Brian

Trabajo Práctico Final

Nota:

- Entregar el TP con nombres y apellidos de los integrantes del grupo.
- **Enviar el TP por mail a snoemiperez@gmail.com** y subir también a "Tareas" de MIEL.
- Presentar este trabajo a modo de informe, comentando la información encontrada. Toda "salida" de software y/o figura que se incluya deberá servir al estudio del problema y deberá ser comentada adecuadamente. Asimismo, las conclusiones y comentarios deben apoyarse en tablas y/o gráficos.
- Envíe el informe y el código que lo respalda, nombrando el archivo con el número de grupo.

Enunciado

Los datos de la base **chicos25.xls** corresponden a niños pequeños a quienes se diagnosticó miopía. El objetivo es modelar esta característica respecto a las variables disponibles para interpretar la relación de algunas variables respecto al diagnóstico.

El detalle de las variables del dataset se encuentran en la segunda hoja del archivo.

Se pide realizar responder los siguientes ítems, pero pueden completar el análisis con estudios adicionales.

1. Analice las variables disponibles para incluirlas de modo adecuado. Estudie faltantes y datos atípicos e indique cómo decide tratarlos. Justifique.

2. Se quiere ajustar modelos de regresión logística que permitan relacionar la miopía con las variables medidas. Para esto, antes que nada, separar los datos en conjuntos de entrenamiento y validación en forma aleatoria en 70/30. Indique que cantidad de casos quedaron para cada ambiente.

3. Considere los siguientes modelos y compárelos adecuadamente en el conjunto de **entrenamiento**:

- Modelo1: un modelo con todas las variables disponibles.
- Modelo2: seleccionando con step sobre todas las variables disponibles.
- Modelo3: un modelo a su elección.

4. Para el modelo elegido en el punto anterior (seguramente ya consideró alguno de estos ítems, se pide aquí que los muestre en el modelo elegido).

a) ¿Son todas las variables significativas?
b) Considere un test de bondad de ajuste: ¿qué conclusión se obtiene?
c) Elija uno de los coeficientes del modelo elegido e interprételo en términos de los odds.
d) ¿Hay problemas de multicolinealidad en las regresoras?
e) ¿Hay datos influyentes? ¿Los mantuvo en el dataset o los quitó?

5. Ajuste un modelo Naive Bayes para predecir la miopía en niños, a partir de las variables disponibles.

6. Finalmente, evalúe y compare **en el conjunto de test** los modelos considerados en los puntos 4 y 5. En particular:
a) Indique AUC y grafique la curva ROC en el conjunto de test.
b) Encuentre la tabla de clasificación y calcule las métricas usuales de comparación.

7. Explique detalladamente cuál sería su elección entre los modelos propuestos en los puntos 4 y 5, para cada una de las situaciones que sean objetivo del estudio:
- Detectar la mayor cantidad de niños con miopía para aplicarles un nuevo tratamiento, teniendo en cuenta que este tratamiento es costoso, por lo que no se quiere aplicar erróneamente a quien no lo necesite.
- Obtener un modelo que haya cometido la menor cantidad de errores de predicción en el proceso de validación.

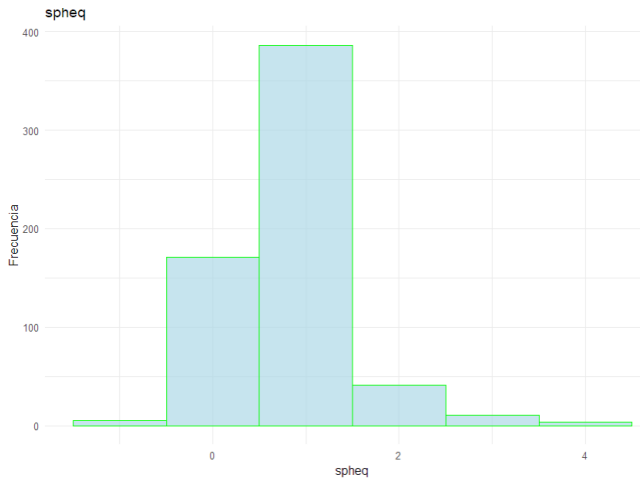
Resolución

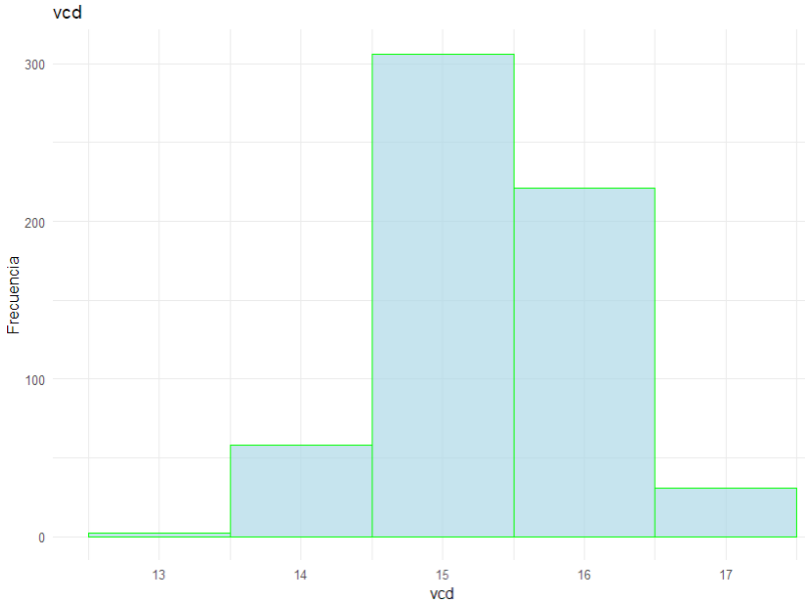
Información del Dataset

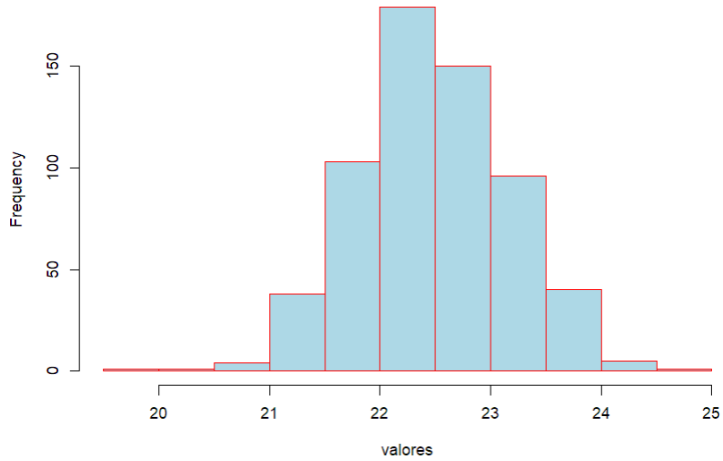
Nombre del dataset: “chicos25.xlsx”
Atributos:
1)” id”: Cuantitativa/numérica (Discreta). Identificador único, no es una variable analítica
2)” mio”: Cualitativa/nominal (Binaria) [Si/No]
3)” spheq”: Cuantitativa/numérica (Continua) [Medidas optométricas]
4)” al”: Cuantitativa/numérica (Continua) [Medidas optométricas]
5)” acd”: Cuantitativa/numérica (Continua) [Medidas optométricas]
6)” vcd”: Cuantitativa/numérica (Continua) [Medidas optométricas]
7)” sporthr”: Cuantitativa/numérica (Continua) [horas dedicadas al deporte en el mes]
8)” tvhr”: Cuantitativa/numérica (Continua) [horas dedicadas a TV en el mes]
9) "ma": Cualitativa/nominal (Binaria) [Si/No] [si la madre tiene miopía]

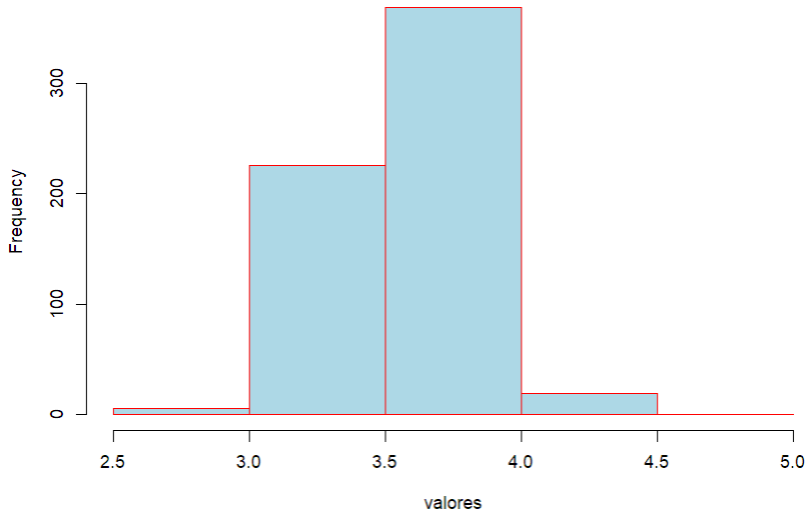
1- Analice las variables disponibles para incluirlas de modo adecuado. Estudie faltantes y datos atípicos e indique cómo decide tratarlos. Justifique.

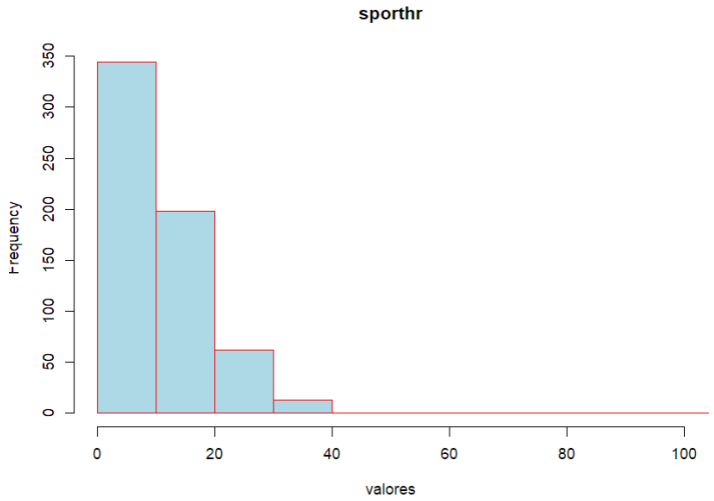
- **Análisis univariado**

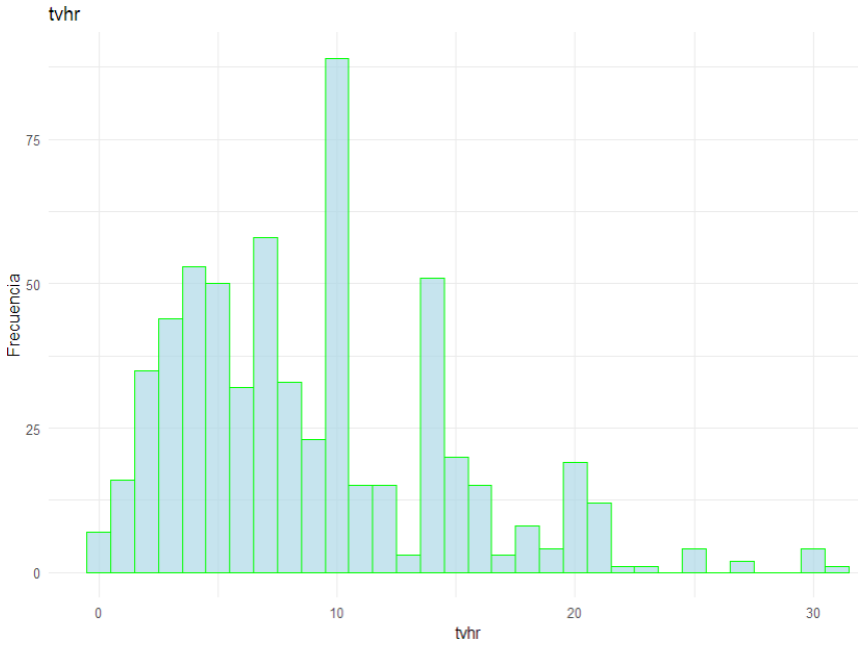



vcd
Frecuencia
300
200
100
0
13
14
15
16
17
vcd


al
Frequency
150
100
50
0
20
21
22
23
24
25
valores


acd
Frequency
300
200
100
0
2.5
3.0
3.5
4.0
4.5
5.0
valores


sporthr
Frequency
0
50
100
150
200
250
300
350
0
20
40
60
80
100
valores


tvhr
Frecuencia
0
25
50
75
0
10
20
30
tvhr

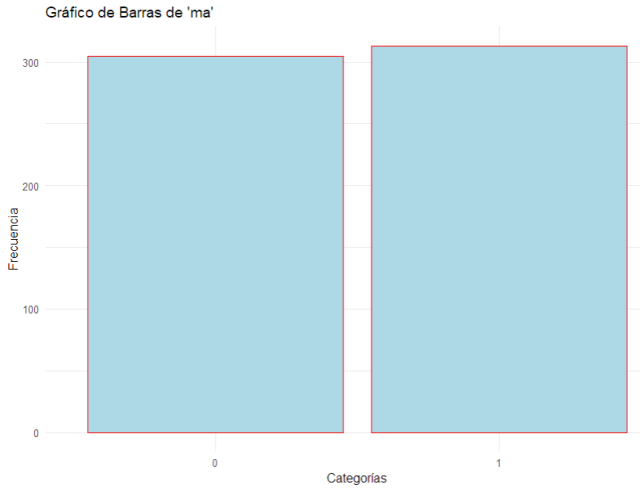


Gráfico de Barras de 'ma'
Frecuencia
0
100
200
300
0
1
Categorías

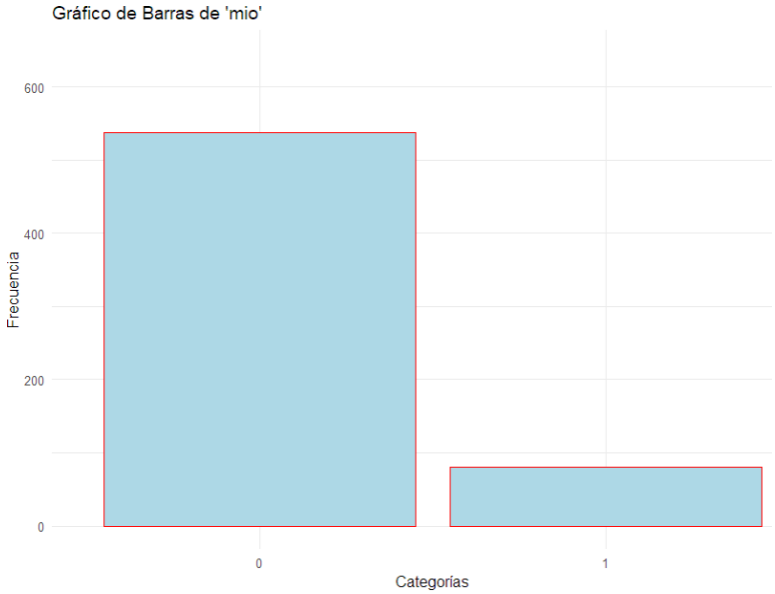
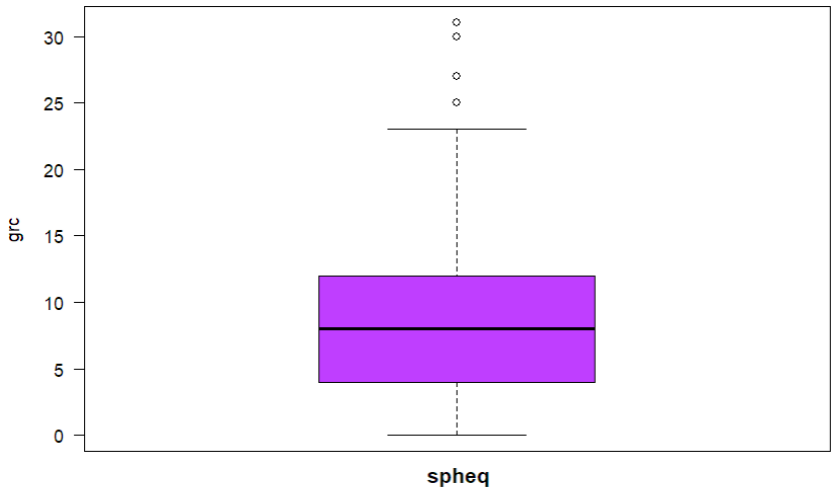
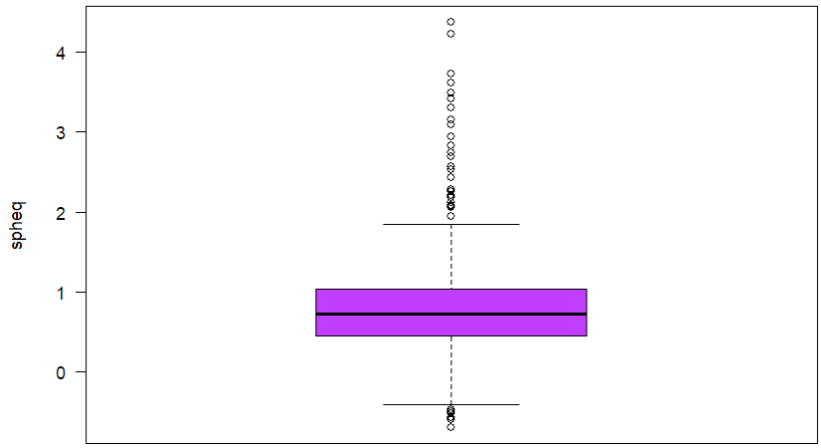
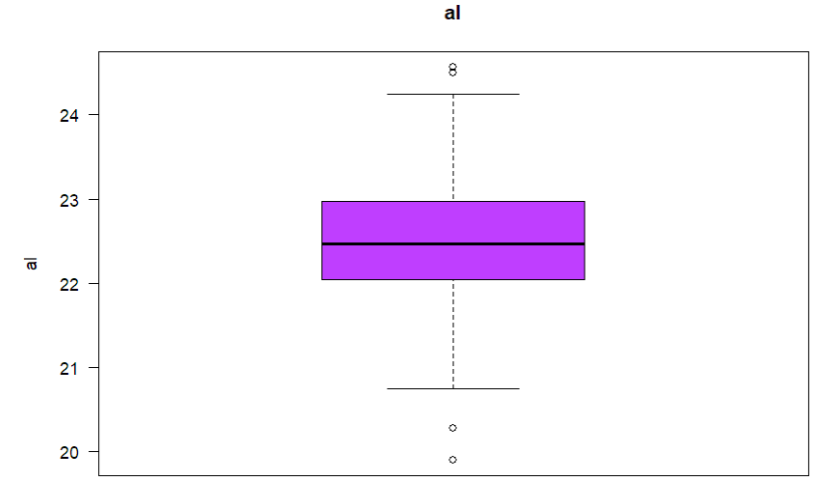


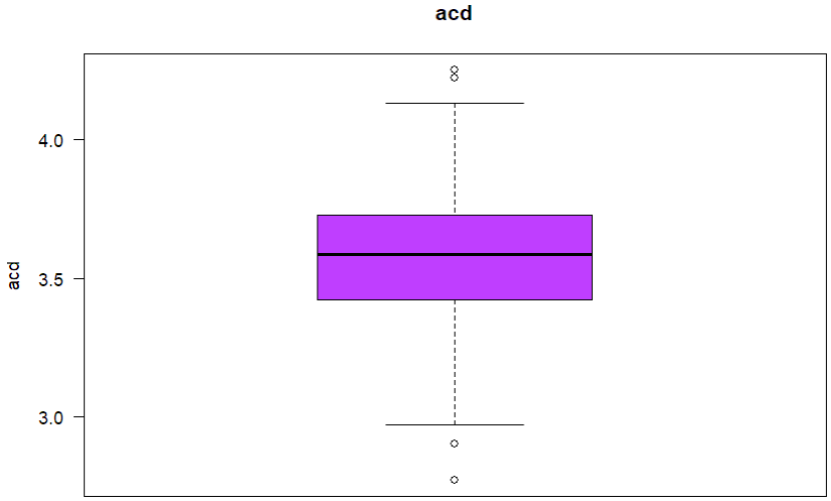
Gráfico de Barras de 'mio'
Frecuencia
600
400
200
0
0
1
Categorías

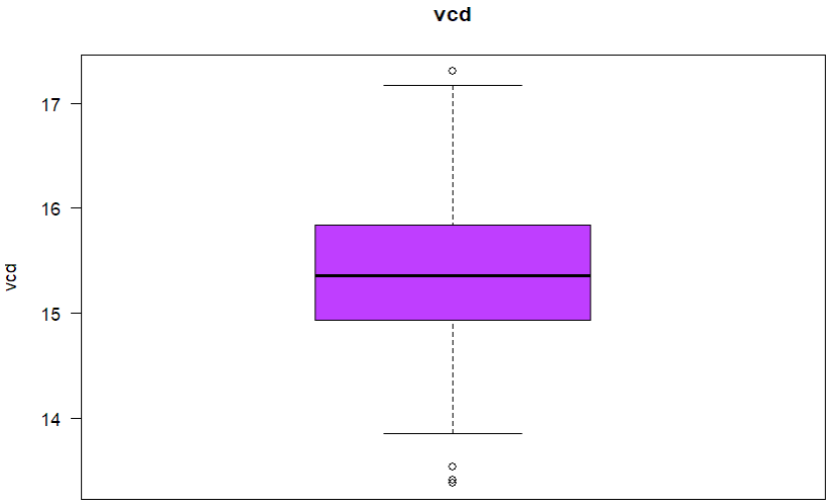

grc
30
25
20
15
10
5
0

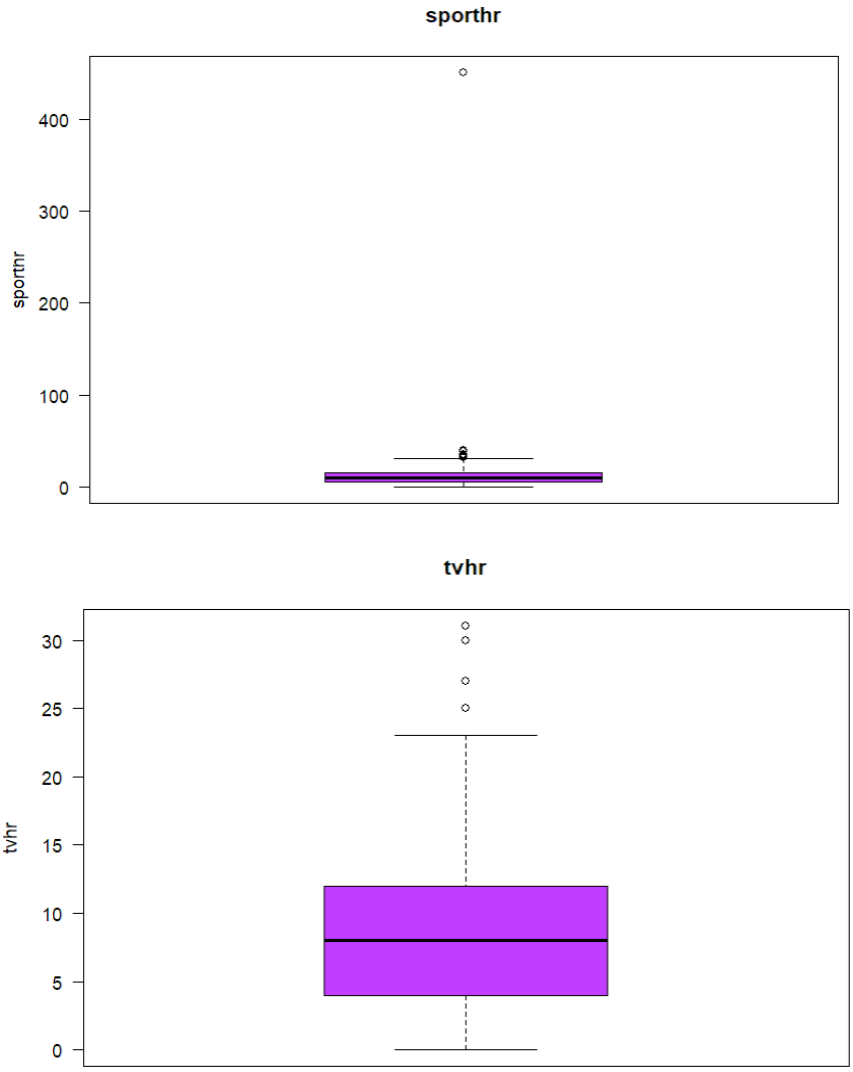
spheq


spheq
4
3
2
1
0


al
24
23
22
21
20
al


acd
4.0
3.5
3.0
acd


vcd
17
16
15
14
vcd



Observaciones:

1)" mio": Presencia de miopía. La variable objetivo mio está desbalanceada 537 casos (86.9%): sin miopía (0) y 81 casos (13.1%) con miopía (1).

0	1
537	81

2)" spheq": Se observa una distribución asimétrica positiva. Los valores negativos indican miopía, mientras que los positivos indican hipermetropía.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.6990000	0.4550000	0.7280000	0.8010081	1.0340000	4.3720000

3)" al": Longitud axial del ojo, en mm. En adultos, la longitud axial típica es aprox. 23–24 mm.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
19.90000	22.04000	22.47000	22.49716	22.97000	24.56000

4)" acd": Profundidad de la cámara anterior, en mm.

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
2.772000 3.424000 3.586000 3.578708 3.730000 4.250000
```

5)" vcd": Profundidad de la cámara vítrea, en mm

```
    Min.  1st Qu.  Median     Mean  3rd Qu.     Max.
13.38000 14.93000 15.36000 15.37699 15.84000 17.30000
```

6)" sporthr": Horas de deporte/mes. Media: 12.52 horas/mes (aprox 3 horas/semana). Máximo 450 horas/mes (aprox. 15 horas/día), outlier con error de registro, se eliminará de nuestro dataset.

```
    Min.  1st Qu.  Median     Mean  3rd Qu.     Max.
 0.00000  6.00000 10.00000 12.52188 16.00000 450.00000
```

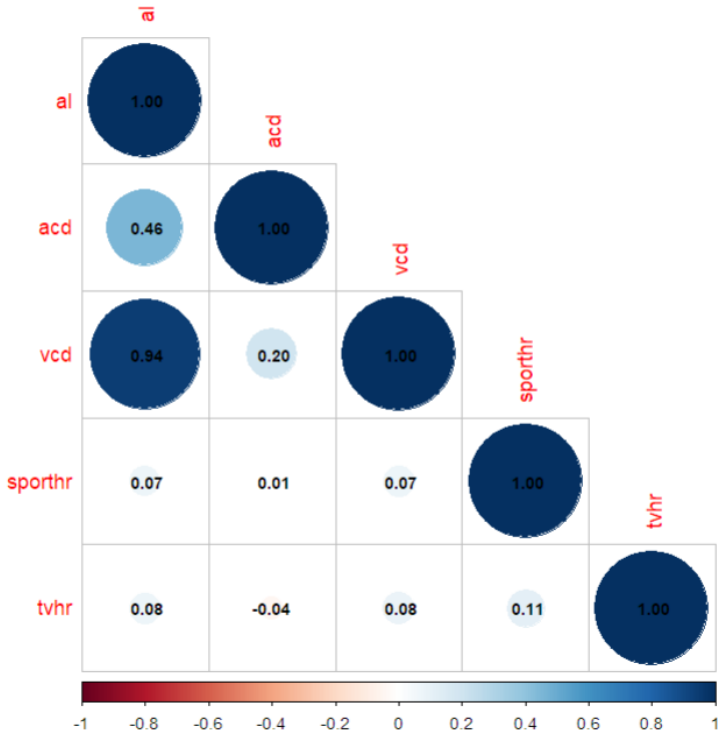
7)" tvhr": Horas de TV/mes. Media: 8.95 horas/mes (aprox 2 horas/semana). Máximo 31 horas/mes, (aprox. 1 hora/día): Valores razonables sin outliers extremos.

```
 Min. 1st Qu.  Median     Mean  3rd Qu.     Max.
 0.0 4.000000 8.000000 8.946515 12.000000 31.000000
```

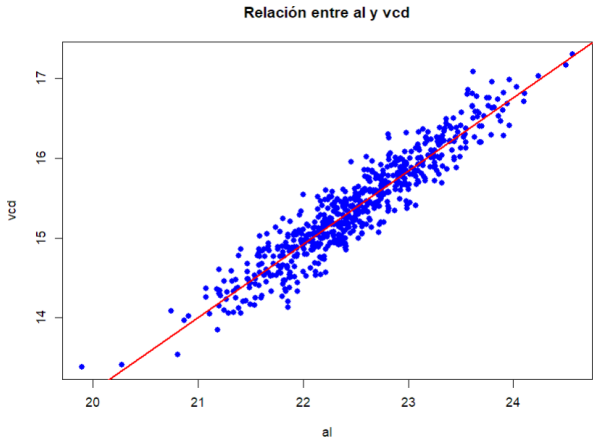
8)" ma": Miopía materna, presenta una distribución equilibrada.

```
  0   1
305 313
```

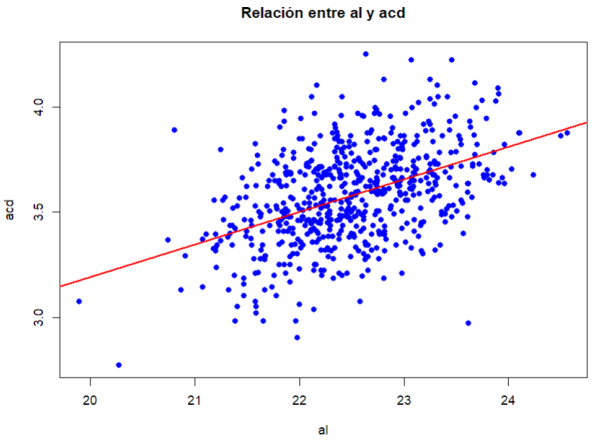
- **Analisis bivariado**



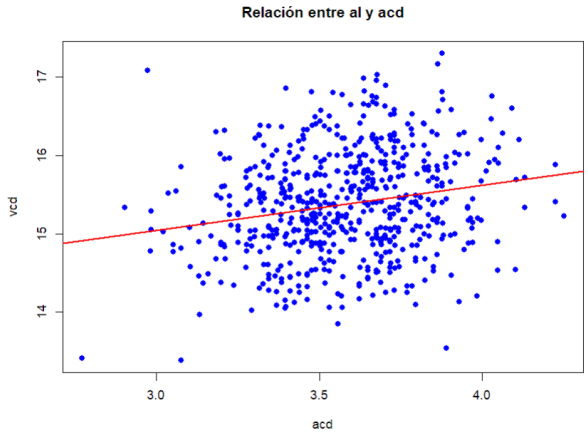
Se destaca la elevada correlación entre las variables al, acd y vcd y la falta de correlación entre sporthr y tvhr con las demás variables.
A continuación, se muestra la relación existente entre las variables más correlacionadas.




```
> coef(modelo)
(Intercept)          al
 -5.3179498   0.9198909
```




```
> coef(modelo)
(Intercept)          al
  0.1016024   0.1545575
```




```
> coef(modelo)
(Intercept)          acd
 13.3198611   0.5748231
```

Se observa la alta relación entre al y vcd, con una recta de regresión con pendiente de 0,920; a medida que al se eleva lo hace también vcd.
Entre al y acd hay una correlación más débil (pendiente de recta de regresión de 0,155), por lo que un aumento en al provoca un pequeño aumento en acd Para el caso de al nulo se tiene que acd alcanza un valor de 0,102.
Por último, entre acd y vcd la recta presenta una menor pendiente en comparación con la recta de al-vcd, pero mayor que la de al-acd.

2- Se quiere ajustar modelos de regresión logística que permitan relacionar la miopía con las variables medidas. Para esto, antes que nada, separar los datos en conjuntos de entrenamiento y validación en forma aleatoria en 70/30. Indique que cantidad de casos quedaron para cada ambiente.

Tras la separación de muestras para entrenamiento y prueba quedaron la siguiente distribución de muestras para cada uno:

```
str(datos) #
summary(datos)
str(datos) # verifico por ultima vez las variables a usar
datos
datos <- datos[, -1] #BORRO 1ER COLUMNA-ID
set.seed(666) #DEFINO SEMILLA ALEATORIA
#attach(datos)
# Cargar el paquete
library(caret)
train<- createDataPartition(mio, p=0.7, list=FALSE) #con list false resultado se devuelve como vector de índice de fila
train #muestras para entrenamiento - son 433 filas/muestras en el vector
nrow(datos) #Nro de muestras de datos
dataTrain <- datos[train, ]
dataTest <- datos[-train, ] #lo restante (30%) para prueba
datos
summary(datos)
```

Cantidad de muestras para entrenamiento: 433
Cantidad de muestras para prueba: 185

3- Considere los siguientes modelos y compárelos adecuadamente en el conjunto de entrenamiento:

- **Modelo1: un modelo con todas las variables disponibles.**
- **Modelo2: seleccionando con step sobre todas las variables disponibles.**
- **Modelo3: un modelo a su elección.**

Comparación de modelos según AIC (Criterio de Información de Akaike):

- Completo

```
> #alternativa 1
> modeloCompleto <- glm(formula = mio ~., data = dataTrain, family = "binomial")
> # me avisaba antes que modelo no convergia! SOLUCION:me olvide de pasar a factor la variable ma!!! ya solucionado
> summary(modeloCompleto)

Call:
glm(formula = mio ~ ., family = "binomial", data = dataTrain)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 10.416115   9.042117   1.152   0.2493
spheq       -3.932435   0.531876  -7.394 1.43e-13 ***
al          -0.699871   1.379270  -0.507   0.6119
acd          1.606404   1.382863   1.162   0.2454
vcd         -0.006203   1.285040  -0.005   0.9961
sporthr     -0.043118   0.024068  -1.792   0.0732 .
tvhr        -0.028733   0.031359  -0.916   0.3595
ma1          0.824433   0.363879   2.266   0.0235 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 337.02  on 431  degrees of freedom
Residual deviance: 227.77  on 424  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 243.77

Number of Fisher Scoring iterations: 7
```

- Automático con step

```
> #alternativa 2
> modeloB <- step(modeloCompleto, direction = "backward") #parto de modelo completo y elimino las var menos significativas
Start:  AIC=243.77
mio ~ spheq + al + acd + vcd + sporthr + tvhr + ma

          Df Deviance    AIC
- vcd      1   227.77 241.77
- al       1   228.03 242.03
- tvhr     1   228.65 242.65
- acd      1   229.12 243.12
<none>         227.77 243.77
- sporthr  1   231.59 245.59
- ma       1   233.15 247.15
- spheq    1   320.79 334.79

Step:  AIC=241.77
mio ~ spheq + al + acd + sporthr + tvhr + ma

          Df Deviance    AIC
- tvhr     1   228.66 240.66
<none>         227.77 241.77
- acd      1   231.18 243.18
- sporthr  1   231.60 243.60
- ma       1   233.17 245.17
- al       1   233.37 245.37
- spheq    1   320.87 332.87

Step:  AIC=240.66
mio ~ spheq + al + acd + sporthr + ma

          Df Deviance    AIC
<none>         228.66 240.66
- acd      1   232.54 242.54
- sporthr  1   233.45 243.45
- ma       1   233.74 243.74
- al       1   235.06 245.06
- spheq    1   320.92 330.92
```

```
> summary(modeloB)

Call:
glm(formula = mio ~ spheq + al + acd + sporthr + ma, family = "binomial",
    data = dataTrain)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 10.77486    6.05493   1.780   0.0752 .
spheq       -3.90493    0.52986  -7.370 1.71e-13 ***
al          -0.74613    0.30343  -2.459   0.0139 *
acd          1.71318    0.87828   1.951   0.0511 .
sporthr     -0.04731    0.02372  -1.995   0.0461 *
ma1          0.79604    0.36114   2.204   0.0275 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 337.02  on 431  degrees of freedom
Residual deviance: 228.66  on 426  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 240.66

Number of Fisher Scoring iterations: 7
```

- Manual: Se consideraron las variables más significativas (spheq, al, sporthr, ma).

Alternativa 1:

```
> #alternativa 3 - a mi eleccion (recordar que el intercepto va por defecto y No es necvesatrio definirla)
> #reviso datos del modelo completo y pruebo con:
> modeloC3 <- glm(formula = mio ~ spheq+acd+sporthr+ma , data = dataTrain, family = "binomial") #parto de m
mino las var menos significativas
> summary(modeloC3)

Call:
glm(formula = mio ~ spheq + acd + sporthr + ma, family = "binomial",
    data = dataTrain)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.53031    2.73808  -0.924   0.3554
spheq       -3.64670    0.50324  -7.246 4.28e-13 ***
acd          0.71571    0.76108   0.940   0.3470
sporthr     -0.04301    0.02302  -1.868   0.0617 .
ma1          0.67907    0.35426   1.917   0.0553 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 337.02  on 431  degrees of freedom
Residual deviance: 235.06  on 427  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 245.06

Number of Fisher Scoring iterations: 6
```

Alternativa 2:

```
> modeloC2 <- glm(formula = mio ~ spheq+acd+ma , data = dataTrain, family = "binomial")
var menos significativas
> summary(modeloC2)

Call:
glm(formula = mio ~ spheq + acd + ma, family = "binomial", data = dataTrain)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.6177     2.7658  -0.946   0.3439
spheq        -3.5925     0.4932  -7.285 3.22e-13 ***
acd           0.5989     0.7644   0.784   0.4333
ma1           0.6897     0.3501   1.970   0.0488 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 337.02  on 431  degrees of freedom
Residual deviance: 239.58  on 428  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 247.58

Number of Fisher Scoring iterations: 6
```

Alternativa 3:

```
> modeloC <- glm(formula = mio ~ spheq+sporthr+ma , data = dataTrain, family = "binomial
las var menos significativas
> summary(modeloC)

Call:
glm(formula = mio ~ spheq + sporthr + ma, family = "binomial",
    data = dataTrain)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.009944   0.426730   0.023   0.9814
spheq       -3.653354   0.501881  -7.279 3.35e-13 ***
sporthr     -0.040537   0.022692  -1.786   0.0740 .
ma1          0.718075   0.352542   2.037   0.0417 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 337.02  on 431  degrees of freedom
Residual deviance: 235.94  on 428  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 243.94

Number of Fisher Scoring iterations: 6
```

Tras comparar los 3 modelos notamos que existe poca diferencia entre los valores AIC y en el caso de modelo automático con step se aprecia una leve diferencia respecto al modelo manual "Alternativa 3" (en adelante hare referencia a esta alternativa únicamente).
Dado que el AIC del modelo completo fue de 243,77, del modelo automático 240,66 y del mejor modelo manual, 243,94 y siendo que en general se busca el AIC mas bajo porque indica que el modelo tiene un buen ajuste con una penalización adecuada por complejidad, a fin de conservar el modelo más parsimonioso, en principio se optaría por el modelo manual, ya que realiza la predicción con solo 3 variables predictoras, mientras que el automático utiliza 5.
A fin de determinar cuál es el modelo que mejor ajusta se realiza un ANOVA, detallado en el siguiente punto.

4- Para el modelo elegido en el punto anterior (seguramente ya consideró alguno de estos ítems, se pide aquí que los muestre en el modelo elegido).

- **¿Son todas las variables significativas?**
- **Considere un test de bondad de ajuste: ¿qué conclusión se obtiene?**
- **Elija uno de los coeficientes del modelo elegido e interprételo en términos de los odds.**
- **¿Hay problemas de multicolinealidad en las regresoras?**
- **¿Hay datos influyentes? ¿Los mantuvo en el dataset o los quitó?**

a- Las variables significativas son: spheq y ma.

```
Call:
glm(formula = mio ~ spheq + sporthr + ma, family = "binomial",
    data = dataTrain)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.009944   0.426730   0.023   0.9814
spheq       -3.653354   0.501881  -7.279 3.35e-13 ***
sporthr     -0.040537   0.022692  -1.786   0.0740 .
ma1          0.718075   0.352542   2.037   0.0417 *
---
```

b- Efectuados en el punto anterior, se detalla el análisis de ANOVA en la siguiente figura

```
Analysis of Deviance Table

Model 1: mio ~ spheq + al + acd + vcd + sporthr + tvhr + ma
Model 2: mio ~ spheq + al + acd + sporthr + ma
Model 3: mio ~ spheq + sporthr + ma
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       424     227.77
2       426     228.66 -2  -0.8841  0.64273
3       428     235.94 -2  -7.2862  0.02617 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Considerando como hipótesis:
H_0: las medias son iguales.
H_1:al menos una media es diferente.

Como la Pr(>Chi) del modelo 2 (automático) dio 0,643, sugiere que el modelo es suficiente y mas simple sin perder poder explicativo significativo, y para el caso del modelo 3 (manual), un Pr(>Chi) de 0,026, por lo la diferencia si es significativa, y el modelo 3 con 3 variables ha perdido capacidad explicativa al eliminar 4 variables del Modelo 1 (completo), por lo que no seria tan adecuado como el modelo 1 o modelo 2.

Por lo tanto, se tiene 2 alternativas:

Opción A: Priorizar Parsimonia (Modelo 3)

- Ventajas: Más simple (3 variables), fácil de interpretar. AIC similar al de modelo completo y levemente superior al modelo automatico.
- Riesgos: Podría subestimar el efecto de variables clínicamente relevantes (al, acd).

Opción B: Priorizar Ajuste (Modelo 2)

- Ventajas: Incluye al y acd, que son relevantes según ANOVA.
- Riesgos: Mayor complejidad (5 variables).

De los resultados obtenidos y priorizando la parsimonia del modelo sobre el ajuste se opta por la opción A, Modelo 3, pero se deberá consultar con un profesional del ámbito clínico si resulta apropiado eliminar del modelo las variables al y acd.

c- Debido a que nuestra variable objetivo es categórica binaria usaremos el modelo de regresión logística.

$$\pi(x) = P(Y = 1/x_1, .. x_p) = \frac{e^{\beta_0 + \beta_1 x_1 + .. \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + .. \beta_p x_p}}$$

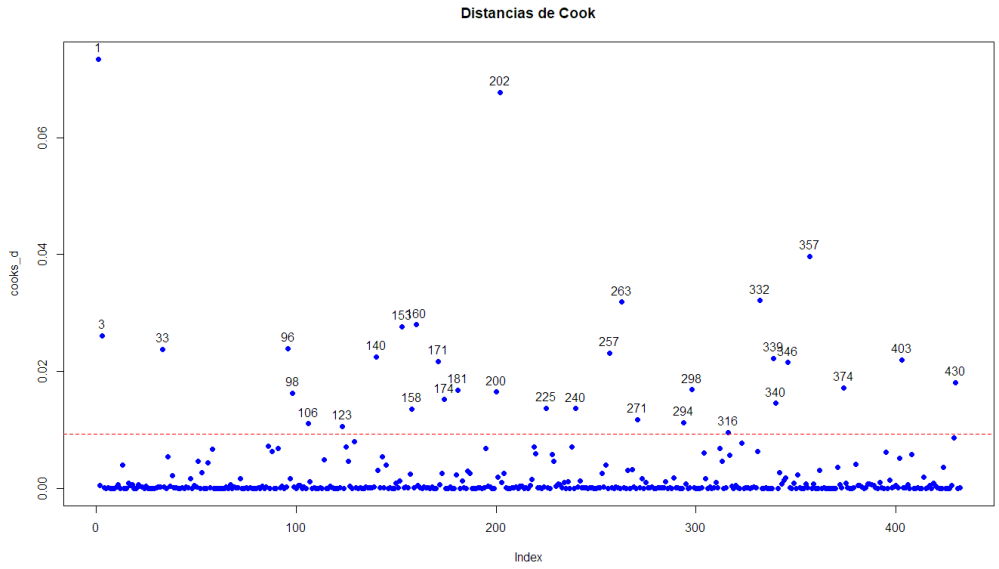
$$\frac{odds(x_i + 1)}{odds(x_i)} = e^{\beta i} = OR(x_i)$$

```
> summary(modeloC)

Call:
glm(formula = mio ~ spheq + sporthr + ma, family = "binomial",
    data = dataTrain)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.009944   0.426730   0.023   0.9814
spheq       -3.653354   0.501881  -7.279 3.35e-13 ***
sporthr     -0.040537   0.022692  -1.786   0.0740 .
ma1          0.718075   0.352542   2.037   0.0417 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 337.02  on 431  degrees of freedom
Residual deviance: 235.94  on 428  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 243.94

Number of Fisher Scoring iterations: 6

> exp(coef(modeloC)["spheq"])
     spheq
0.02590409
> exp(coef(modeloC)["ma1"])
     ma1
2.050481
```

$$\text{Odds Ratio} = \exp(\text{Coeficiente de spheq}) = \exp(-3{,}653354) = 0{,}026 \text{ (aprox)}$$

Por cada aumento de una unidad en la variable spheq, el odds de que la respuesta sea "SI" (presencia de miopía) se reduce en un factor de aproximadamente 0,026. En otras palabras, un aumento en spheq está asociado con una disminución en las “chances” de que un individuo tenga miopía, ya que el odds ratio es menor que 1.
Respecto al coeficiente de ma sugiere que cuando la variable ma está presente (es decir, ma = 1 con miopía materna), las probabilidades de que un individuo tenga miopía aumentan en un factor de aproximadamente 2.05 en comparación con cuando ma está ausente (es decir, ma = 0 sin miopía materna).

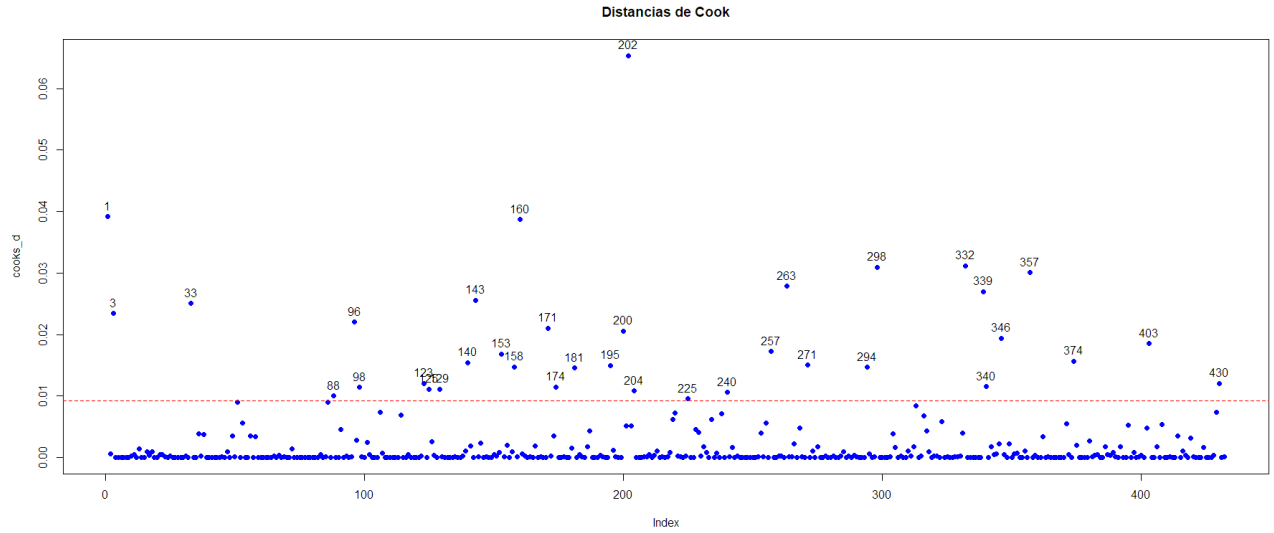
d- A fin de tratar la multicolinealidad entre las variables regresoras se procede a calcular la distancia de Cook y se aprecian outliers. Se considera el umbral de 4/n (n=nro de muestras de train, 433) en línea punteada roja.



e- En función del grafico de distancias de Cook, se observan valores influyentes a revisar su consideración para el análisis posterior. Al desconocer el dominio que pueden alcanzar estas variables se opta por mantener estas muestras.

```
> print(datos[obs_influyentes, ], n = Inf)
# A tibble: 32 x 8
   mio    spheq    al   acd   vcd sporthr  tvhr ma
   <fct>  <dbl> <dbl> <dbl> <dbl>   <dbl> <dbl> <fct>
 1 1     -0.052  21.9  3.69  14.7      40    10 1
 2 0      1.18   22.5  3.46  15.5      14    10 0
 3 0      1.09   23.0  3.62  15.9      14    14 1
 4 0     -0.051  21.7  3.13  14.9      10    11 1
 5 1     -0.215  23.0  3.84  15.7       3     3 1
 6 0      0.807  22.9  3.49  16.0      10    10 0
 7 1     -0.6    22.7  3.61  15.6       1    10 1
 8 0      1.03   23.0  3.82  15.8       5     7 0
 9 0      0.765  23.5  3.50  16.6      35    10 1
10 0      0.553  22.9  3.90  15.7      10     8 1
11 0      0.777  23.0  3.97  15.4      10    15 1
12 0      0.621  23.1  3.82  15.9       6     3 1
13 0      0.697  21.2  3.34  14.3      15    10 0
14 0      0.684  23.4  3.86  15.9      10     2 0
15 0      0.89   21.4  3.44  14.2      12    10 1
16 0      3.49   22.2  4.10  14.5      10     3 1
17 0      0.789  22.3  3.72  15.0       6     4 0
18 0      1.68   22.7  3.40  15.8      20    10 0
19 0      1.36   21.8  3.41  14.3      15     6 1
20 0      0.777  22.2  3.72  14.8      10     5 1
21 0      1.15   23.0  3.70  15.9      10     6 0
22 0      0.72   21.8  3.52  14.6       4     4 0
23 0      0.434  21.9  3.42  14.9       9     9 0
24 0      1.01   22.0  3.65  14.4      12    11 1
25 0      0.308  22.9  3.61  15.8      10     3 1
26 0      0.193  21.7  3.28  14.6      13     9 0
27 0      1.32   22.4  3.40  15.4       6    12 0
28 1      1.32   22.3  3.33  15.3       3     7 1
29 0      0.334  22.1  3.45  14.9      10     5 1
30 0      1.03   21.8  3.77  14.3      15    10 1
31 0      0.285  23.7  3.82  16.4      15    30 1
32 0      0.815  23.0  3.60  15.8      14    18 0
```

Para el modelo automático se obtiene la siguiente distribución de distancias de Cook:




```
> print(datos[obs_influyentes, ], n = Inf)
# A tibble: 36 × 8
   mio    spheq    al   acd   vcd sporthr  tvhr ma
   <fct>  <dbl> <dbl> <dbl> <dbl>   <dbl> <dbl> <fct>
 1 1     -0.052  21.9  3.69  14.7      40    10 1
 2 0      1.18   22.5  3.46  15.5      14    10 0
 3 0      1.09   23.0  3.62  15.9      14    14 1
 4 0      1.13   23.1  3.64  15.8      10     5 1
 5 0     -0.051  21.7  3.13  14.9      10    11 1
 6 1     -0.215  23.0  3.84  15.7       3     3 1
 7 1     -0.6    22.7  3.61  15.6       1    10 1
 8 0      0.138  21.6  3.41  14.4      14    17 0
 9 0      0.567  22.4  3.96  14.9       6     4 1
10 0      1.03   23.0  3.82  15.8       5     7 0
11 1      0.234  21.5  3.41  14.4      10    10 1
12 0      0.765  23.5  3.50  16.6      35    10 1
13 0      0.553  22.9  3.90  15.7      10     8 1
14 0      0.777  23.0  3.97  15.4      10    15 1
15 0      0.621  23.1  3.82  15.9       6     3 1
16 0      0.697  21.2  3.34  14.3      15    10 0
17 0      0.684  23.4  3.86  15.9      10     2 0
18 0      0.933  21.4  2.98  14.8       5     5 1
19 0      0.89   21.4  3.44  14.2      12    10 1
20 0      3.49   22.2  4.10  14.5      10     3 1
21 0      0.488  21.8  3.62  14.4      26    14 0
22 0      0.789  22.3  3.72  15.0       6     4 0
23 0      1.68   22.7  3.40  15.8      20    10 0
24 0      1.36   21.8  3.41  14.3      15     6 1
25 0      0.777  22.2  3.72  14.8      10     5 1
26 0      1.15   23.0  3.70  15.9      10     6 0
27 0      0.72   21.8  3.52  14.6       4     4 0
28 0      0.434  21.9  3.42  14.9       9     9 0
29 0      0.308  22.9  3.61  15.8      10     3 1
30 0      0.193  21.7  3.28  14.6      13     9 0
31 0      1.32   22.4  3.40  15.4       6    12 0
32 1      1.32   22.3  3.33  15.3       3     7 1
33 0      0.334  22.1  3.45  14.9      10     5 1
34 0      1.03   21.8  3.77  14.3      15    10 1
35 0      0.285  23.7  3.82  16.4      15    30 1
36 0      0.815  23.0  3.60  15.8      14    18 0
```

Se aprecian más outliers en modelo automático de valores similares que en el modelo manual, validando nuestra selección por el modelo manual.

5- Ajuste un modelo Naive Bayes para predecir la miopía en niños, a partir de las variables disponibles.

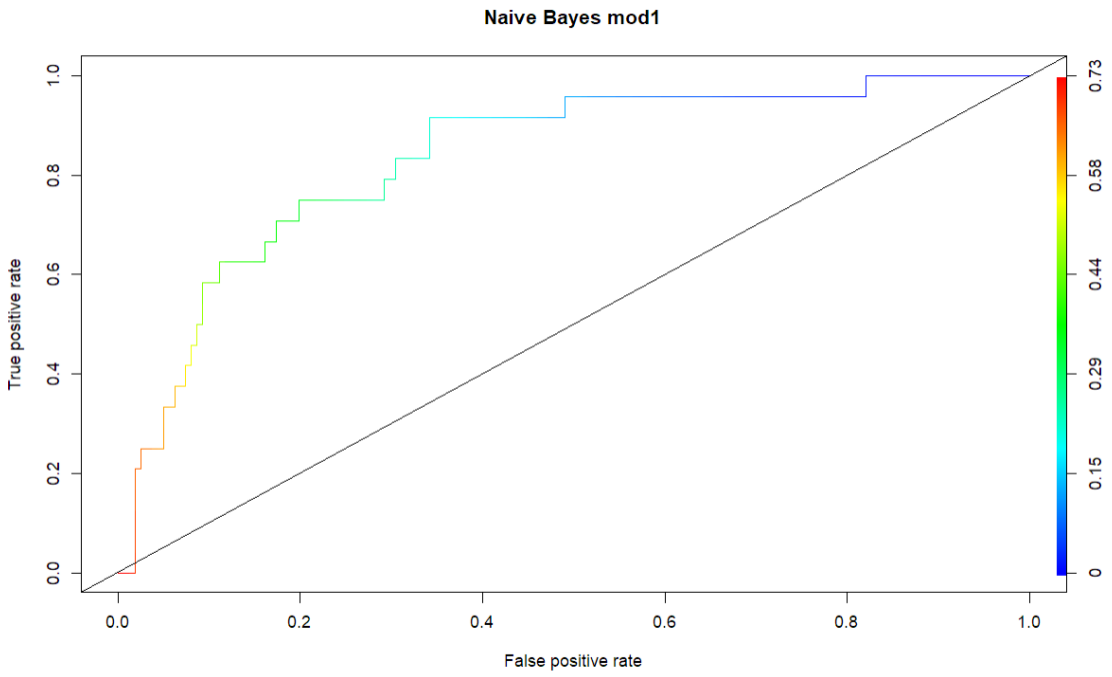
A partir de las variables del modelo manual se desarrolla el modelo de Naive Bayes.

```
> dataTest[1,] #el caso 1 del conjunto de Test
# A tibble: 1 x 8
  mio   spheq    al   acd   vcd sporthr  tvhr ma
  <fct> <dbl> <dbl> <dbl> <dbl>   <dbl> <dbl> <fct>
1 0      1.18  22.5  3.46  15.5      14    10 0
> predict(object = mod1, newdata=dataTest[1,], type = "raw") #probabilidades predichas
          0         1
[1,] 0.9833328 0.01666719
> predict(object = mod1, newdata=dataTest[1,], type = "class") #clase predicha
[1] 0
Levels: 0 1
> dataTest[2,]
# A tibble: 1 x 8
  mio   spheq    al   acd   vcd sporthr  tvhr ma
  <fct> <dbl> <dbl> <dbl> <dbl>   <dbl> <dbl> <fct>
1 0     0.987  22.9  3.70  15.9      14     4 0
> predict(object = mod1, newdata=dataTest[2,], type = "raw") #probabilidades predichas
          0         1
[1,] 0.9516407 0.04835928
> dataTest[51,]
# A tibble: 1 x 8
  mio   spheq    al   acd   vcd sporthr  tvhr ma
  <fct> <dbl> <dbl> <dbl> <dbl>   <dbl> <dbl> <fct>
1 0     0.147  23.6  3.60  16.6      17    10 0
> predict(object = mod1, newdata=dataTest[51,], type = "raw") #probabilidades predichas
          0         1
[1,] 0.6144864 0.3855136
> dataTest # verifico que la prediccion da igual a lo que tengo en mi muestra de prueba -- el modelo predice ok
# A tibble: 185 x 8
   mio   spheq    al   acd   vcd sporthr  tvhr ma
   <fct> <dbl> <dbl> <dbl> <dbl>   <dbl> <dbl> <fct>
 1 0      1.18  22.5  3.46  15.5      14    10 0
 2 0     0.987  22.9  3.70  15.9      14     4 0
 3 0     0.169  23.0  3.50  15.9      10    20 1
 4 0     0.466  24.0  3.70  16.9       8     3 0
 5 0      1.03  23.3  3.73  16.3       4     3 0
 6 0      1.40  24.0  3.82  16.4      25    10 0
 7 1      0.49  23.1  3.58  16         10    12 1
 8 1      0.67  22.6  3.65  15.6        5     5 1
 9 0     0.071  21.7  3.14  15.1        2    14 1
10 0      1.09  23.0  3.62  15.9      14    14 1
```

6- Finalmente, evalúe y compare en el conjunto de test los modelos considerados en los puntos 4 y 5. En particular:

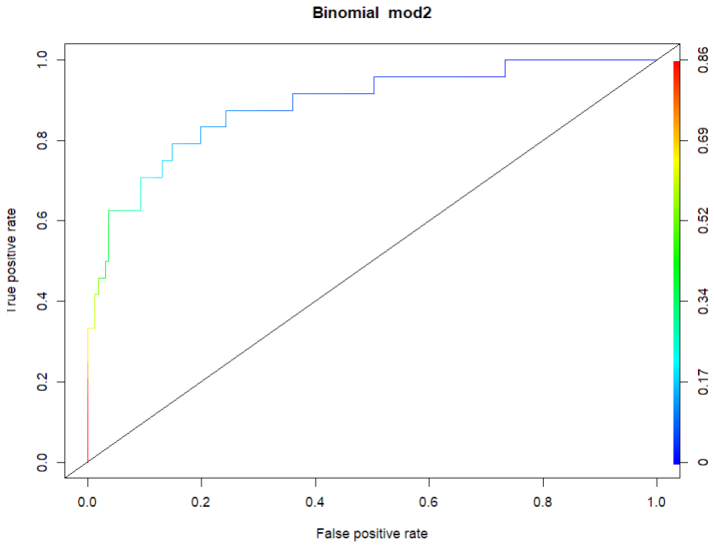
- **Indique AUC y grafique la curva ROC en el conjunto de test.**
- **Encuentre la tabla de clasificación y calcule las métricas usuales de comparación.**

Curva ROC en el conjunto de test para modelo de Bayes:



```
> #CURVA ROC Y AUC
> library(ROCR)
> prediccion1<-prediction(proba.1[,2],dataTest$mio)
> roc_mod1 <- performance(prediccion1, measure = "tpr", x.measure = "fpr")
> plot(roc_mod1, main = "Naive Bayes mod1", colorize = T)
> abline(a = 0, b = 1)
> AUC.mod1 <- performance(prediccion1, "auc")
> #para que me de AUC:
> AUC.mod1@y.values
[[1]]
[1] 0.8356625
```

Curva ROC en el conjunto de test para modelo manual Binomial:




```
> prediccion2<-prediction(proba.2[,],dataTest$mio)# traigo todas los elementos de la columna de proba.2
lo tenia unindice y valor)
> roc_mod2 <- performance(prediccion2, measure = "tpr", x.measure = "fpr")
> plot(roc_mod2, main = "Binomial  mod2", colorize = T)
> abline(a = 0, b = 1)
> AUC.mod2 <- performance(prediccion2, "auc")
> #para que me de AUC:
> AUC.mod2@y.values
[[1]]
[1] 0.88794
```

En términos de la interpretación de la curva ROC, mientras más cerca de la esquina superior izquierda esté la curva, mejor será el rendimiento del modelo, ya que esto indica que hay una alta tasa de verdaderos positivos con una baja tasa de falsos positivos.

Comparando la curva del modelo binomial y la de Bayes, la que posee mayor rendimiento es la binomial y presenta mayor AUC.

b- A continuación, se aprecia la tabla de clasificación y las métricas usuales de comparación.

Bayes:

```
> confusionMatrix(confusion.1,positive = "1")
Confusion Matrix and Statistics

          predicho
observado   0   1
        0 147  14
        1  13  11

               Accuracy : 0.8541
                 95% CI : (0.7948, 0.9016)
    No Information Rate : 0.8649
    P-Value [Acc > NIR] : 0.7109

                  Kappa : 0.3649

 Mcnemar's Test P-Value : 1.0000

            Sensitivity : 0.44000
            Specificity : 0.91875
         Pos Pred Value : 0.45833
         Neg Pred Value : 0.91304
             Prevalence : 0.13514
         Detection Rate : 0.05946
   Detection Prevalence : 0.12973
      Balanced Accuracy : 0.67937

       'Positive' Class : 1
```

Binomial

```
> confusion.2 <- table(dataTest$mio, predi.2,dnn = c("observado","predicho"))
> confusionMatrix(confusion.2,positive = "1")
Confusion Matrix and Statistics

          predicho
observado   0   1
        0 158   3
        1  14  10

               Accuracy : 0.9081
                 95% CI : (0.857, 0.9456)
    No Information Rate : 0.9297
    P-Value [Acc > NIR] : 0.89848

                  Kappa : 0.4945

 Mcnemar's Test P-Value : 0.01529

            Sensitivity : 0.76923
            Specificity : 0.91860
         Pos Pred Value : 0.41667
         Neg Pred Value : 0.98137
             Prevalence : 0.07027
         Detection Rate : 0.05405
   Detection Prevalence : 0.12973
      Balanced Accuracy : 0.84392

       'Positive' Class : 1
```

- Exactitud (Accuracy): El modelo binomial tiene una mejor exactitud 0,908 en comparación con el modelo Bayes que tiene 0,854.

- Sensibilidad (Recall o Tasa de Verdaderos Positivos): el modelo de binomial presenta mayor valor, de 0,769 frente a los 0,44 de Bayes, para identificar correctamente los casos positivos
- Especificidad (Tasa de Verdaderos Negativos): valores similares para ambos modelos de 0,919 para identificar correctamente los casos negativos
- Kappa: El modelo binomial tiene un valor de Kappa superior al del modelo de Bayes, 0,495 frente a 0,365, para el acuerdo entre las predicciones del modelo y las verdaderas clases
- Balanced Accuracy: El modelo binomial también supera al modelo Bayes, 0,844 frente a 0,679, lo que sugiere un mejor rendimiento general ((sensibilidad + especificidad)/2).

El modelo binomial supera al modelo Bayes en las métricas claves, por lo que tiene un mejor desempeño en la clasificación general.

7- Explique detalladamente cuál sería su elección entre los modelos propuestos en los puntos 4 y 5, para cada una de las situaciones que sean objetivo del estudio:

- **Detectar la mayor cantidad de niños con miopía para aplicarles un nuevo tratamiento, teniendo en cuenta que este tratamiento es costoso, por lo que no se quiere aplicar erróneamente a quien no lo necesite.**
- **Obtener un modelo que haya cometido la menor cantidad de errores de predicción en el proceso de validación.**

Los datos de semilla utilizados son 161 casos de no miopes y 24 de miopes.

```
> summary(dataTest)
 mio         spheq              al              acd             vcd            sporthr          tvhr            ma
 0:161   Min.   :-0.502   Min.   :20.75   Min.   :3.022   Min.   :13.85   Min.   : 0.00   Min.   : 0.000   0:87
 1: 24   1st Qu.: 0.487   1st Qu.:22.12   1st Qu.:3.462   1st Qu.:14.94   1st Qu.: 6.00   1st Qu.: 5.000   1:98
         Median : 0.738   Median :22.53   Median :3.622   Median :15.41   Median :10.00   Median : 9.000
         Mean   : 0.789   Mean   :22.55   Mean   :3.599   Mean   :15.42   Mean   :11.99   Mean   : 9.027
         3rd Qu.: 1.044   3rd Qu.:22.97   3rd Qu.:3.730   3rd Qu.:15.85   3rd Qu.:17.00   3rd Qu.:12.000
         Max.   : 3.731   Max.   :24.11   Max.   :4.090   Max.   :16.96   Max.   :31.00   Max.   :25.000
>
> #----------------------------------------7----------------------------------------
> summary(datos)
 mio         spheq              al              acd             vcd            sporthr          tvhr            ma
 0:536   Min.   :-0.699   Min.   :19.90   Min.   :2.772   Min.   :13.38   Min.   :  0.00   Min.   : 0.000   0:305
 1: 81   1st Qu.: 0.455   1st Qu.:22.04   1st Qu.:3.424   1st Qu.:14.93   1st Qu.:  6.00   1st Qu.: 4.000   1:312
         Median : 0.728   Median :22.47   Median :3.586   Median :15.36   Median : 10.00   Median : 8.000
         Mean   : 0.801   Mean   :22.50   Mean   :3.579   Mean   :15.38   Mean   : 12.52   Mean   : 8.947
         3rd Qu.: 1.034   3rd Qu.:22.97   3rd Qu.:3.730   3rd Qu.:15.84   3rd Qu.: 16.00   3rd Qu.:12.000
         Max.   : 4.372   Max.   :24.56   Max.   :4.250   Max.   :17.30   Max.   :450.00   Max.   :31.000
```

a-

A fin de detectar la mayor cantidad de niños con miopía para aplicarles un nuevo tratamiento, se opta por realizar la tabla de matriz de confusión a fin de no aplicar erróneamente a quien no lo necesite.

Matriz de confusion de Bayes:

```
> confusionMatrix(confusion.1,positive = "1")
Confusion Matrix and Statistics

          predicho
 VN   FP  0   1
             7  14
 FN   VP  3  11
```

Matriz de confusion para modelo binomial con probabilidad umbral de 0,5 (si es mayor es miope):

```
> confusionMatrix(confusion.2,positive = "1")
Confusion Matrix and Statistics

          predicho
VN    FP  0   1
FN    VP  3   3
          4  10
```

Dado que el tratamiento no solo debe realizarse a quienes lo necesitan, sino también evitarse en quienes no lo requieren debido a su alto costo, se considera reducir la cantidad de falsos positivos (FP) y aumentar los casos de verdaderos positivos (VP), por lo que se considera la métrica Pos Pred Value (Precision).

$$Precision = \frac{VP}{VP + FP}$$

La precisión del modelo de Bayes es de 0,44 y del modelo binomial de 0,77, por lo que la opción más conveniente es el modelo binomial.

Dado que el tratamiento no solo debe realizarse a quienes lo necesitan, sino también evitarse en quienes no lo requieren debido a su alto costo, se considera reducir la cantidad de falsos positivos (FP), y el modelo binomial lo logra con 3 casos frente a los 14 del modelo de Bayes.

b- El modelo que comete la menor cantidad de errores de predicción es el que tiene la menor cantidad de falsos positivos y falsos negativos, resultando para nuestros modelos:
Bayes: 14+13=27
Binomial: 3+14=17

Por lo que el modelo con menor cantidad de errores es el Binomial.

Mejoras futuras

¿Qué pasaría si se varía la probabilidad de umbral?

Para el caso de p umbral de 0,3 se tiene un aumento en la cantidad de verdaderos positivos y un leve aumento en las predicciones erróneas (FP + FN), respecto al modelo binomial con p umbral de 0,5.

	predicho	
observado	0	1
0	152	9
1	9	15

En caso de optar por p umbral de 0,4, se tiene un leve aumento de predicciones verdaderas positivas manteniendo la cantidad de predicciones erróneas (FP+FN), respecto al modelo binomial con p umbral de 0,5.

	predicho	
observado	0	1
0	156	5
1	12	12

En caso de optar por p umbral de 0,1, se duplican las predicciones verdaderas positivas, pero se eleva las predicciones erróneas (FP+FN), respecto al modelo binomial con p umbral de 0,5.

	predicho	
observado	0	1
0	119	42
1	3	21

Por último, de considerarse un p umbral de 0,8, se reduce a menos de la mitad las predicciones verdaderas positivas y se elevan las predicciones erróneas (FP+FN), respecto al modelo binomial con p umbral de 0,5.

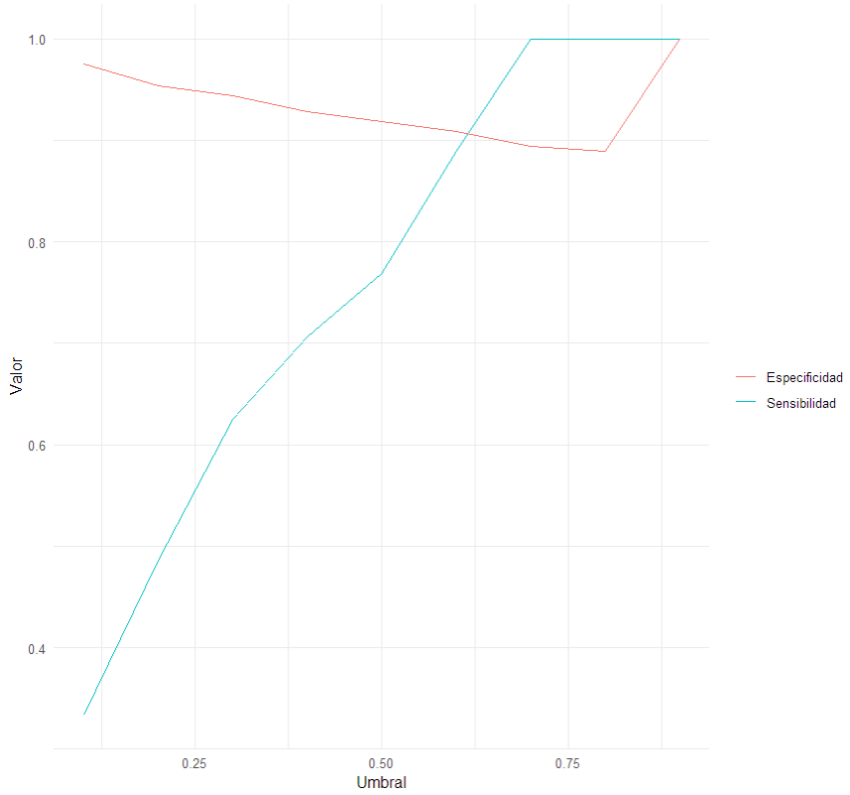
	predicho	
observado	0	1
0	161	0
1	20	4

Se observa como al aumentar la p umbral en el modelo binomial se reducen los FP y VP. Los FP disminuyen porque el modelo ya no clasifica casos negativos como positivos fácilmente, y algunos casos de VP no alcanzaran el umbral.
Se aprecia como una p umbral de 0,4 y 0,5 alcanzan un equilibrio entre predicciones acertadas (VP) y falsos positivos (FP).

En la siguiente figura se aprecia diferentes métricas para una iteración con pasos de 0,1 para la p umbral.

```
> df_mod.log
  Umbral  Accuracy Precision Sensibilidad Especificidad       F1
1    0.1 0.7567568 0.8750000    0.3333333     0.9754098 0.4827586
2    0.2 0.8648649 0.7083333    0.4857143     0.9533333 0.5762712
3    0.3 0.9027027 0.6250000    0.6250000     0.9440994 0.6250000
4    0.4 0.9081081 0.5000000    0.7058824     0.9285714 0.5853659
5    0.5 0.9081081 0.4166667    0.7692308     0.9186047 0.5405405
6    0.6 0.9081081 0.3333333    0.8888889     0.9090909 0.4848485
7    0.7 0.8972973 0.2083333    1.0000000     0.8944444 0.3448276
8    0.8 0.8918919 0.1666667    1.0000000     0.8895028 0.2857143
9    0.9 1.0000000 1.0000000    1.0000000     1.0000000 1.0000000
```

Se verifican los resultados obtenidos tras compararlos con los parámetros calculados anteriormente de forma manual para p umbral de 0,5.
En la siguiente curva se aprecia la variación de sensibilidad (verdaderos positivos) y la especificidad (verdaderos negativos) al variar la p umbral.



Se aprecia como una p umbral de 0,5 es próxima a un p umbral óptimo entre estos parámetros. Además de la tabla se observa la variación de F1 (combinación de precisión y sensibilidad) y se observa cómo se eleva al bajar el p valor.
De esta forma una p umbral del orden de 0,5 alcanza un equilibrio entre las diferentes métricas.

Alternativa 2 en modelo binomial

Si se considera la alternativa 2, donde se pondera el ajuste en base al análisis ANOVA y AIC, y se descarta la parsimonia del ajuste al modelo completo, los siguientes parámetros se ven afectados.

Coeficientes del modelo binomial

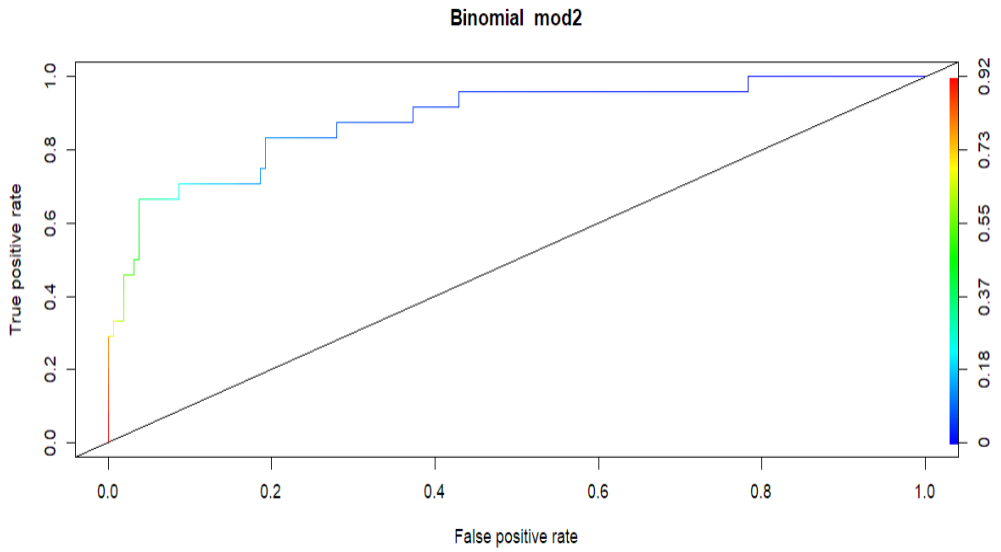
El modelo binomial de 5 variables ajusta al modelo completo para predecir mio tienen los siguientes coeficientes para las 5 variables predictoras:

```
> coef(summary(modeloB))
               Estimate Std. Error   z value     Pr(>|z|)
(Intercept) 10.77486135 6.05493016  1.779519 7.515476e-02
spheq       -3.90492996 0.52985927 -7.369749 1.709492e-13
al          -0.74612660 0.30342626 -2.459005 1.393228e-02
acd          1.71317655 0.87828433  1.950594 5.110531e-02
sporthr     -0.04731313 0.02372087 -1.994578 4.608891e-02
ma1          0.79604402 0.36114327  2.204233 2.750794e-02
```

Se aprecia que todas las variables son significativas a diferencia de la alternativa 3 (modelo manual C) considerado anteriormente.

Curva ROC y AUC

Curva ROC en el conjunto de test para el modelo:




```
> AUC.mod2 <- performance(prediccion2, "auc")
> #para que me de AUC:
> AUC.mod2@y.values
[[1]]
[1] 0.8848344
```

Recordando que mientras más cerca de la esquina superior izquierda esté la curva, mejor será el rendimiento del modelo, se alcanza un valor de AUC de 0,885 mayor al del modelo de Bayes considerado anteriormente, y prácticamente igual al obtenido con la alternativa 3 (manual) del modelo binomial, de 0,888.

Tabla de clasificación y las métricas usuales de comparación

```
Confusion Matrix and Statistics

          predicho
observado   0   1
        0 157   4
        1  13  11

               Accuracy : 0.9081
                 95% CI : (0.857, 0.9456)
    No Information Rate : 0.9189
    P-Value [Acc > NIR] : 0.75620

                  Kappa : 0.5158

 Mcnemar's Test P-Value : 0.05235

            Sensitivity : 0.73333
            Specificity : 0.92353
         Pos Pred Value : 0.45833
         Neg Pred Value : 0.97516
             Prevalence : 0.08108
         Detection Rate : 0.05946
   Detection Prevalence : 0.12973
      Balanced Accuracy : 0.82843

       'Positive' Class : 1
```

```
Confusion Matrix and Statistics

          predicho
observado   0   1
        0 158   3
        1  14  10

               Accuracy : 0.9081
                 95% CI : (0.857, 0.9456)
    No Information Rate : 0.9297
    P-Value [Acc > NIR] : 0.89848

                  Kappa : 0.4945

 Mcnemar's Test P-Value : 0.01529

            Sensitivity : 0.76923
            Specificity : 0.91860
         Pos Pred Value : 0.41667
         Neg Pred Value : 0.98137
             Prevalence : 0.07027
         Detection Rate : 0.05405
   Detection Prevalence : 0.12973
      Balanced Accuracy : 0.84392

       'Positive' Class : 1
```

La tabla resaltada en verde corresponde a las métricas del modelo binomial automático y la resaltada en rojo a la del modelo binomial manual.
Se verifica que la cantidad de casos de observado=1, es decir casos FN y VP, corresponde con la cantidad de muestras de prueba/test (24).
Si se pretende identificar correctamente los positivos (sensibilidad) el modelo manual es levemente más conveniente, pero si se prefiere identificar los negativos (especificidad) el modelo automático presenta un leve mejor rendimiento.
Ambos modelos presentan una exactitud similar, aunque el automático presenta un leve mejor Kappa (concordancia predicción-observado/real).
Se observa que ambos modelos alcanzan la misma cantidad de predicciones erróneas y respecto a la precisión, el modelo automático presenta un rendimiento de 0,73, menor al alcanzado por el modelo manual de 0,77.

Ambos modelos alcanzan métricas prácticamente similares, por lo que priorizar la simplicidad en el modelo, la parsimonia, resulta adecuado para este caso, aun cuando herramientas que evalúan la calidad de los modelos estadísticos, como AIC (Criterio de Información de Akaike) y ANOVA (Análisis de Varianza), prefieran uno más complejo.