Considera los datos del fichero `infartos.RData`, sobre enfermedades coronarias en Sudáfrica. Se obtuvieron en una encuesta llevada a cabo en el marco del Coronary Risk-Factor Study (CORIS) en tres zonas rurales de Western Cape (Sudáfrica).[1] El objetivo del estudio era establecer la intensidad de factores de riesgo para enfermedades coronarias en esa región de alta incidencia.
Los individuos eran varones de raza blanca con edades entre 15 y 64 años. La etiqueta `clase` consignaba la presencia (`clase`=1) o ausencia (`clase`=0) de infarto de miocardio en el momento de la encuesta. Las características medidas fueron:

Nombre variable	Descripción
`sbp`	Tensión sanguínea sistólica
`tobacco`	Consumo de tabaco
`ldl`	Colesterol
`adiposity`	Medida de adiposidad
`typea`	Comportamiento "tipo A"
`obesity`	Medida de la obesidad
`alcohol`	Consumo de alcohol
`age`	Edad

Calcula la función lineal discriminante de Fisher para clasificar entre sano (`clase=0`) o enfermo (`clase=1`) a un individuo en función de las 8 variables regresoras contenidas el fichero. Compara los coeficientes de las variables con los correspondientes a la regla de clasificación basada en regresión logística. ¿Son muy diferentes?

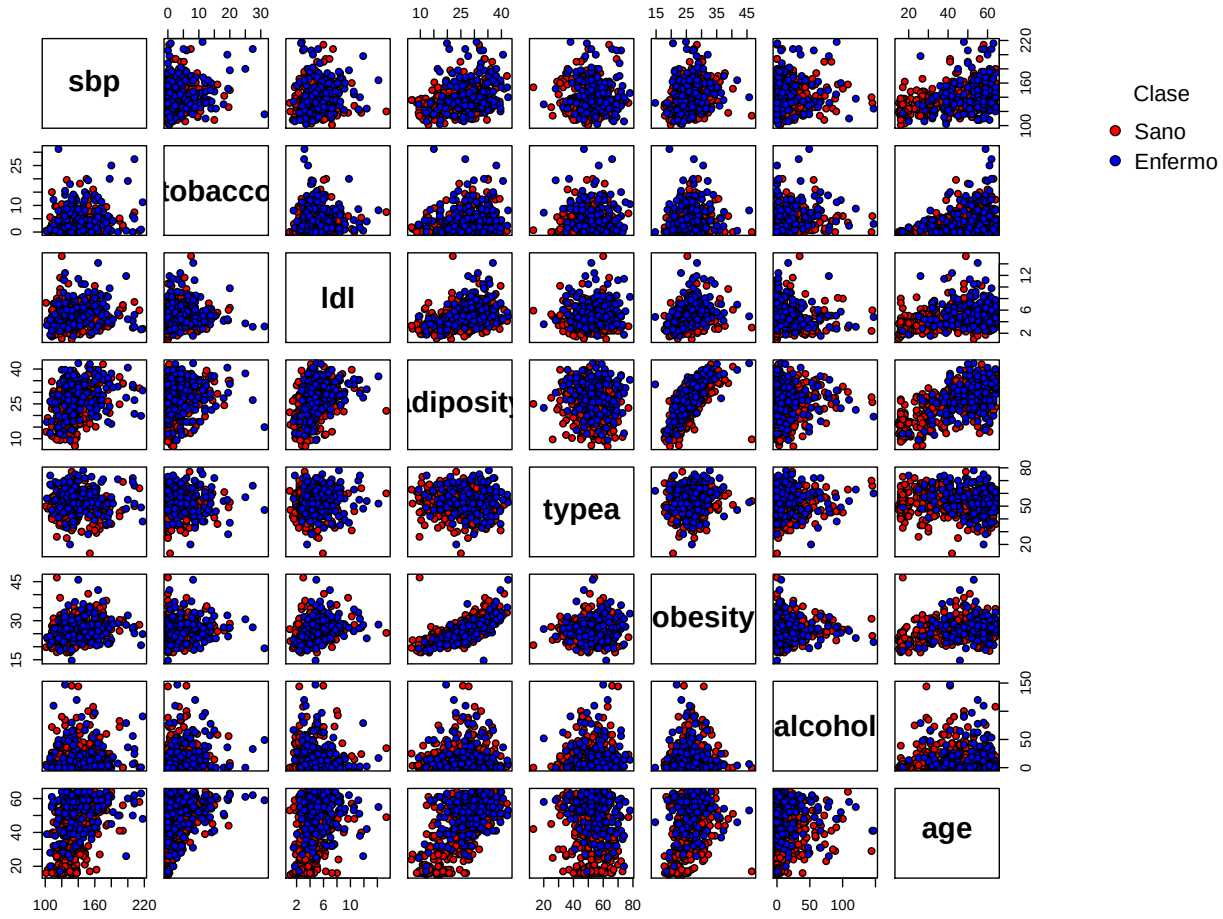
Solución:

```
load("infartos.RData")
# Las caracteristicas se cargan en la tabla datos y la respuesta en el vector
    clase

# Diagrama de dispersion multiple

pdf("InfartosDispMult.pdf",width=10,height=7)
pairs(datos, cex = 1, pch = 21,
      bg=c("red","blue")[unclass(factor(clase))],
      cex.labels = 2, font.labels = 2,oma = c(2,2,2,18))
par(xpd=TRUE)
legend("topright",inset=c(0,0),c("Sano","Enfermo"),pch = 21,
       pt.bg=c("red","blue"),cex=1, text.font=1, title="Clase",bty="n")
dev.off()
```

[1] Rousseauw, J., du Plessis, J., Benade, A., Jordaan, P., Kotze, J., Jooste, P., Ferreira, J. (1983). Coronary risk factor screening in three rural communities, *South African Medical Journal* 64, 430–436.



En el diagrama de dispersión multiple observamos que es un problema de clasificación "complicado", en el sentido de que las dos poblaciones no están claramente separadas. Esto se reflejará debajo en la tasa de error de clasificación del procedimiento de Fisher aproximada mediante validación cruzada: sale cerca de un 30 %.

```
# Regla lineal de Fisher

library(MASS)
ClasFisher = lda(datos,clase)
summary(ClasFisher)

ClasFisher$scaling # Direccion de proyeccion
                  LD1
sbp       6.496208e-03
tobacco   8.203075e-02
ldl       1.920521e-01
adiposity 1.136053e-02
typea     3.414582e-02
obesity  -5.522155e-02
alcohol   1.349809e-05
age       4.260489e-02

ClasFisher$means # Medias por clases
      sbp  tobacco     ldl adiposity  typea obesity alcohol     age
0 135.4603 2.634735 4.344238 23.96911 52.36755 25.73745 15.93136 38.85430
1 143.7375 5.524875 5.487938 28.12025 54.49375 26.62294 19.14525 50.29375

ClasFisherCV = lda(datos,clase,CV=TRUE) # Con CV=T, obtenemos leave-one-out
n = nrow(datos) # Tama\~{n}o muestral
sum(clase != ClasFisherCV$class)/n # TEVC (Tasa de error calculada con leave-one-
    out)
[1] 0.2922078

# Regresion logistica
reglog = glm(clase ~ datos$sbp + datos$tobacco + datos$ldl + datos$adiposity +
    datos$typea + datos$obesity + datos$alcohol + datos$age,family="binomial")
summary(reglog)
Call:
glm(formula = clase ~ datos$sbp + datos$tobacco + datos$ldl +
    datos$adiposity + datos$typea + datos$obesity + datos$alcohol +
    datos$age, family = "binomial")

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-2.0519 -0.8392 -0.4681  0.9825  2.4535

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -6.066864  1.271443 -4.772 1.83e-06 ***
datos$sbp      0.005641  0.005611  1.005 0.314721
datos$tobacco  0.072716  0.026326  2.762 0.005742 **
datos$ldl      0.192492  0.059429  3.239 0.001199 **
datos$adiposity 0.017066 0.028433  0.600 0.548355
datos$typea    0.040467  0.012078  3.350 0.000807 ***
datos$obesity -0.057931  0.042980 -1.348 0.177703
datos$alcohol  0.001446  0.004403  0.328 0.742627
datos$age      0.050650  0.011766  4.305 1.67e-05 ***
```

```
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1       1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11 on 461 degrees of freedom
Residual deviance: 488.89 on 453 degrees of freedom
AIC: 506.89

Number of Fisher Scoring iterations: 4
```

A continuación representamos gráficamente los coeficientes de la regla lineal de Fisher y los de la regla de clasificación obtenida con regresión logística. Veremos que son muy parecidos.

```
pdf("InfartosPesosFeatures.pdf",width=10,height=6)
plot(seq(1,8),ClasFisher$scaling,pch=24,cex=2,axes=F,col="black",bg="grey",
     xlab="Caracteristica",ylab="Peso de caracteristica")
lines(seq(1,8),ClasFisher$scaling,col="grey")
points(seq(1,8),reglog$coefficients[2:9],pch=25,cex=2,col="black",bg="blue")
lines(seq(1,8),reglog$coefficients[2:9],col="blue",lty=2)
axis(1,at=seq(1,8),labels=names(datos))
axis(2)
legend("topright",inset=c(0,0),c("Fisher","Logit"),pch = c(24,25),
       pt.bg=c("grey","blue"),cex=1, text.font=1, title="Clasificador",bty="n")
dev.off()
```