Tema 0: Introducción

Amparo Baíllo Moreno

Departamento de Matemáticas
Universidad Autónoma de Madrid

Datos de contacto

- Amparo Baíllo Moreno
- Despacho: 304 del módulo 17
- Correo electrónico: `amparo.baillo@uam.es`
- Web: `http://verso.mat.uam.es/~amparo.baillo/`

Tutorías (presenciales o en línea): Sólo con cita previa.

Prerrequisitos para Estadística II: Estadística I (contrastes, lema de Fisher, distribuciones normal, χ^2, t-Student y F). Álgebra lineal. Probabilidad I.

Nota final = max (Nota examen final , $(1-p)$ Nota examen final $+ \; p$ Nota parcial) , donde $0 \leq p \leq 0.3$.

Nota mínima en el examen final para hacer media con el parcial: 3.5.

Observaciones

- Tanto en el examen final como en el control podrá utilizarse una calculadora, las **tablas** estadísticas y un formulario **elaborado personalmente por cada alumno en una cara** de una hoja de tamaño DIN A4.

- Las transparencias son material de apoyo para dar la clase. **No son apuntes de la asignatura**.

- Es muy importante **pensar los problemas antes** de que comentemos las soluciones en clase.

- Uno de los objetivos de la asignatura es aprender a analizar datos con el ordenador. **Vamos a utilizar** `R`.

Temario

1. **La distribución normal multivariante.**
   Repaso de vectores aleatorios. Distribuciones marginales y condicionadas bajo normalidad. Independencia. Distribución de formas cuadráticas.
2. **Contrastes no paramétricos.**
   Contrastes de bondad de ajuste. Contrastes de homogeneidad. Contrastes de independencia.
3. **El modelo de regresión lineal.**
   Hipótesis, estimadores de mínimos cuadrados, propiedades de los estimadores, descomposición de la variabilidad, contrastes. Variables regresoras cualitativas. Diagnóstico del modelo y análisis de los residuos.
4. **Clasificación y regresión logística.**
   En función del tiempo disponible trataremos algunos de los siguientes temas: planteamiento del problema de clasificación supervisada, regla de clasificación lineal de Fisher, regresión logística, optimalidad y regla Bayes,...

Bibliografía

Draper, N. and Smith, H. (1998). *Applied Regression Analysis* (3ª ed.) Wiley.

Faraway, J. (2002). *Practical Regression and Anova using R*. `https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf`

Izenman, A.J. (2008). *Modern Multivariate Statistical Techniques*. Springer
Caps. 3, 5 y 8.

Peña, D. (2010). *Regresión y diseño de experimentos* (2ª ed.) Alianza Editorial.

Rice, J.A. (1995). *Mathematical Statistics and Data Analysis* (2ª ed.) Duxbury Press. Caps. 9, 12, 13 y 14.

Weisberg, S. (2005). *Applied Linear Regression* (3ª ed.) Wiley.

Ejemplo 1: Extraído de www.physicsforums.com

Aug26-10, 02:00 PM #1

schip666!

schip666! is Offline:

Posts: 211

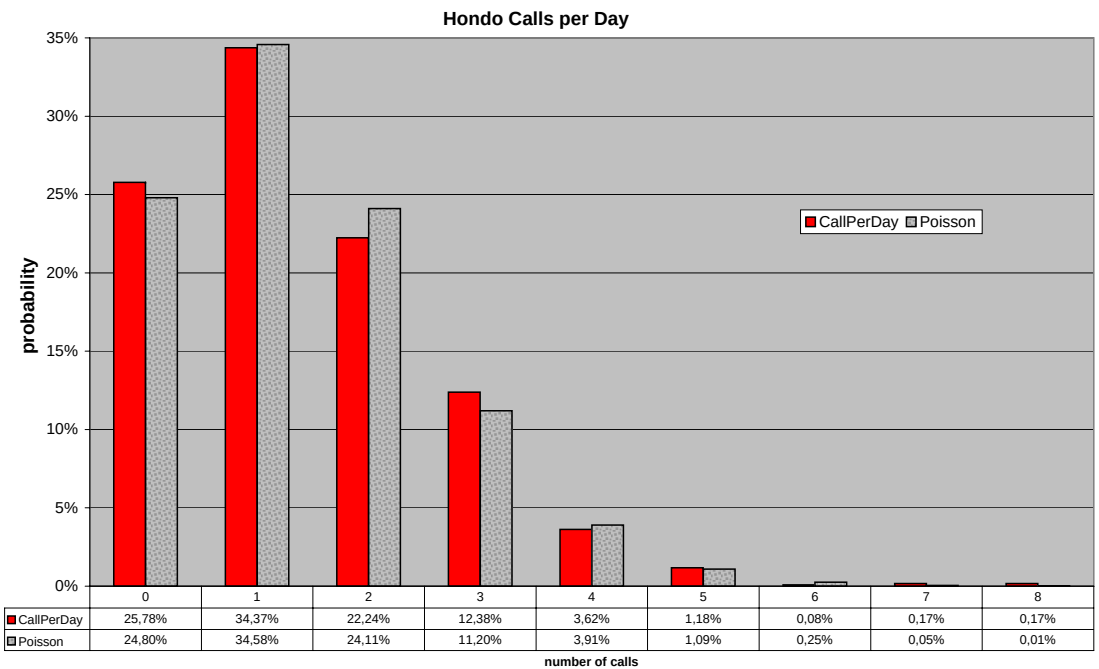
probability of event occurring -- poisson distribution?

I am the keeper of records for my local Volunteer Fire Dept. I have now collected data for each of our incident calls from the last 3 years and have made some _very_ basic stabs at interesting statistics which you can see at:
http://hondovfd.org/statistics.php

We have about 500 calls a year -- a bit over 40 a month or around 1.3 per day. But as you can see from the graphs at the bottom of that page -- which are just about the full extent of my Excel skills -- they are not randomly distributed over the days of the week or hours of the day. More interesting to all our responders is how they are distributed by number per day. My "Calls per Day" graph seems to show a sorta-exponential decay from 1 per day to 8 (our all time high during a snow storm when our little section of Interstate turned into a Bumper Car arena). However we can go for up to a week with nada, and then break the drought with 3 or 4 in an afternoon.

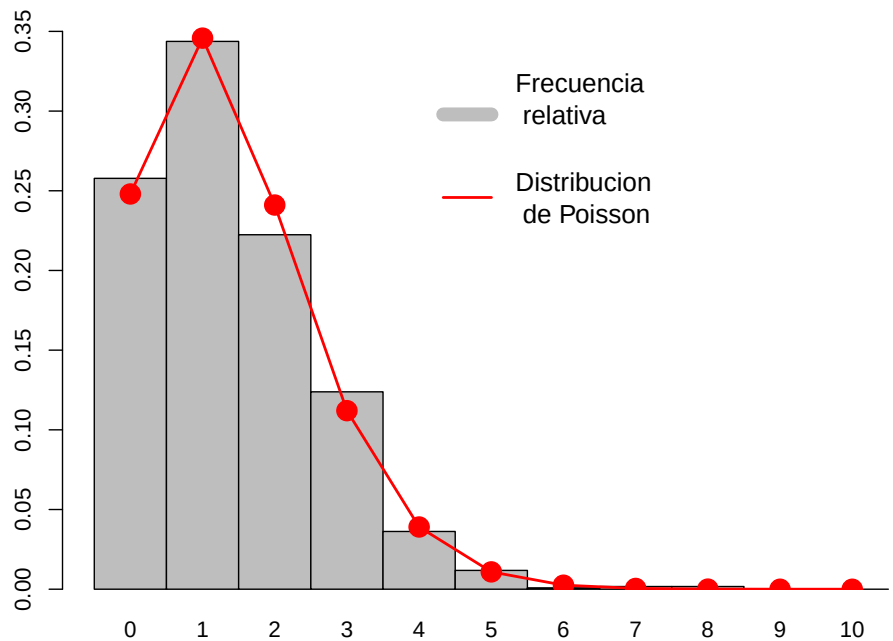
So the question is: How do I characterize the likely-hood of getting a certain number of calls in any particular day, with the number 0 being of special interest. I think I should be able to compare to a Poisson distribution to see how un-random things are, but my eyes roll into the back of my head about a quarter of the way through the wiki page. Can anyone point me to some other explanations and examples, or have better thoughts on the approach?

Nº llamadas diarias	Recuento de días
0	306
1	408
2	264
3	147
4	43
5	14
6	1
7	2
8	2
9	0
10	0
Total	1187 días



	0	1	2	3	4	5	6	7	8
CallPerDay	25,78%	34,37%	22,24%	12,38%	3,62%	1,18%	0,08%	0,17%	0,17%
Poisson	24,80%	34,58%	24,11%	11,20%	3,91%	1,09%	0,25%	0,05%	0,01%

```
Valores = seq(0,10)
Frec = c(306,408,264,147,43,14,1,2,2,0,0)
n = sum(Frec)
m = sum(Valores*Frec)/n
barplot(Frec/n,names=Valores,space=0,ylim=c(0,0.37))
MasaPoisson = dpois(Valores,m)
lines(Valores+0.5,MasaPoisson,type="l",col="red",lwd=2)
points(Valores+0.5,MasaPoisson,pch=19,col="red",cex=2)
legend(x=4.5,y=0.35,c(paste("Frecuencia\n relativa\n"),
                      paste("Distribucion\n de Poisson")),
       col=c("grey","red"),cex=1.2, bty="n",lty=c(1,1),
       lwd=c(10,2),text.font=1)
```



Ejemplo 2: renta y fracaso escolar

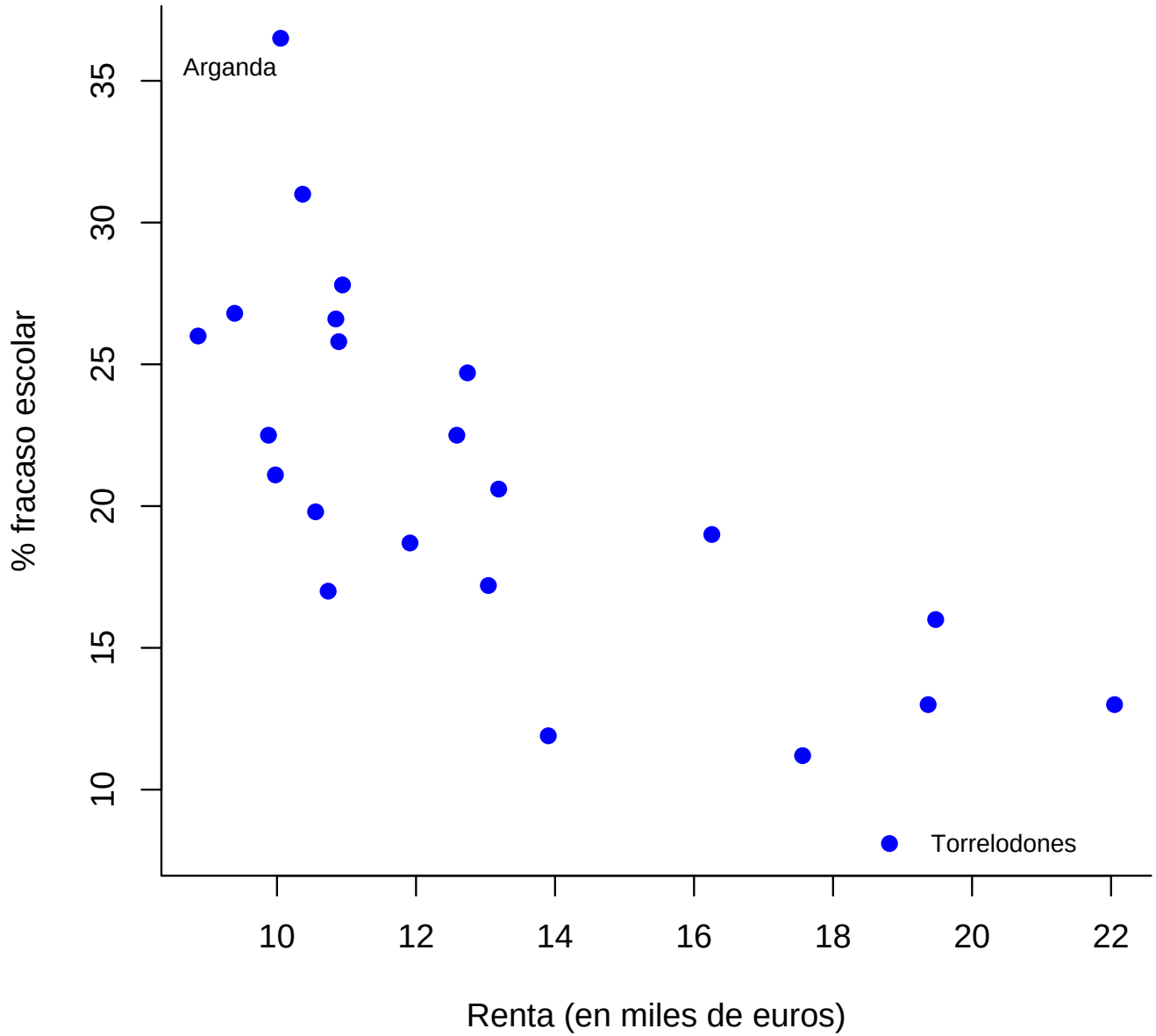
EL PAÍS, martes 18 de octubre de 2005

El fracaso escolar es más alto en las zonas con menor renta

Fracaso escolar en la Comunidad de Madrid

Renta per capita bruta media en 2003: 13.095 euros

	CURSO 2003/2004 Renta (euros)	Fracaso escolar (%)
Parla	8.864	26,0
Fuenlabrada	9.391	26,8
Leganés	9.877	22,5
Móstoles	9.977	21,1
Arganda	10.052	36,5
Torrejón	10.369	31,0
Getafe	10.555	19,8
Coslada	10.736	17,0
Pinto	10.846	26,6
Alcorcón	10.888	25,8
Alcalá de Henarés	10.942	27,8
Collado	11.913	18,7
Colmenar Viejo	12.587	22,5
Arroyomolinos	12.740	24,7
S. Sebastián de los Reyes	13.041	17,2
S. Lorenzo del Escorial	13.189	20,6
Rivas	13.903	11,9
Alcobendas	16.256	19,0
Tres Cantos	17.562	11,2
Torrelodones	18.812	8,1
Boadilla	19.368	13,0
Majadahonda	19.477	16,0
Pozuelo	22.050	13,0


Arganda
Torrelodones
% fracaso escolar
10
15
20
25
30
35
10
12
14
16
18
20
22
Renta (en miles de euros)

```
Datos = read.table("RentaFracaso.txt",header=T)
Datos$Renta= Datos$Renta/1000
X = Datos[,c(2,3)]
plot(X,pch=21,cex=1.5,col="blue",bg="blue",
     xlab="Renta (en miles de euros)",ylab="% fracaso escolar")
text(10,35,"Arganda")
```

Algunas preguntas sobre estos datos

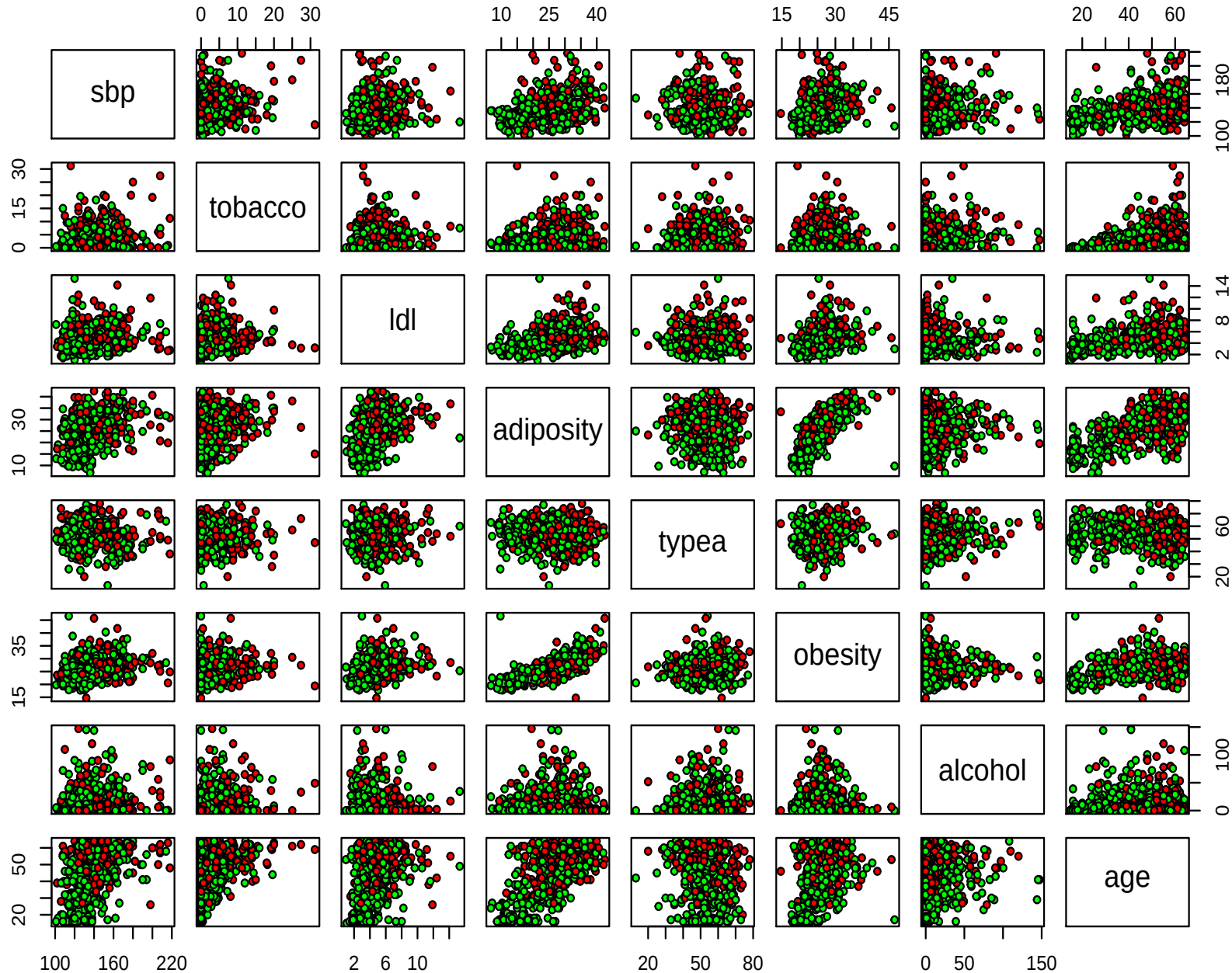
- Para cada una de las variables, ¿podemos suponer que proceden de algún modelo probabilístico concreto? Por ejemplo, podemos suponer que la renta sigue una distribución normal?
- ¿Podemos afirmar que existe relación lineal entre las variables?
- ¿Qué fracaso escolar podemos predecir en una población cuya renta es de 14000 euros?
- ¿Hasta qué punto son fiables las predicciones que hemos obtenido?
- ¿Existen datos atípicos? ¿Cómo influyen los datos atípicos en los resultados?

Ejemplo 3: infartos de miocardio

South African Heart Disease data de Hastie *et al.* (2009).
En un estudio de factores de riesgo en enfermedades coronarias, se dispone de datos de 462 personas (de las que 160 habían sufrido infartos y 302 eran controles). Para cada una de ellas se midieron las siguientes variables:

Nombre variable	Descripción
`sbp`	Tensión sanguínea sistólica
`tobacco`	Consumo de tabaco
`ldl`	Colesterol
`adiposity`	Medida de adiposidad
`famhist`	Historial familiar de enfermedad coronaria. 2 niveles: `Present`/`Absent`
`typea`	Comportamiento "tipo A"
`obesity`	Medida de la obesidad
`alcohol`	Consumo de alcohol
`age`	Edad

Representación gráfica de los datos: verde (controles) y rojo (casos).


sbp
tobacco
ldl
adiposity
typea
obesity
alcohol
age

```
library(readxl)
SAheart <- read_excel("SAheart.xls")

str(SAheart)
 tibble[,11] [462 x 11] (S3: tbl_df/tbl/data.frame)
 $ row      : num [1:462] 1 2 3 4 5 6 7 8 9 10 ...
 $ sbp      : num [1:462] 160 144 118 170 134 132 142 114 114 132 ...
 $ tobacco  : num [1:462] 12 0.01 0.08 7.5 13.6 6.2 4.05 4.08 0 0 ...
 $ ldl      : num [1:462] 5.73 4.41 3.48 6.41 3.5 6.47 3.38 4.59 3.83
    5.8 ...
 $ adiposity: num [1:462] 23.1 28.6 32.3 38 27.8 ...
 $ famhist  : chr [1:462] "Present" "Absent" "Present" "Present" ...
 $ typea    : num [1:462] 49 55 52 51 60 62 59 62 49 69 ...
 $ obesity  : num [1:462] 25.3 28.9 29.1 32 26 ...
 $ alcohol  : num [1:462] 97.2 2.06 3.81 24.26 57.34 ...
 $ age      : num [1:462] 52 63 46 58 49 45 38 58 29 53 ...
 $ chd      : num [1:462] 1 1 0 1 1 0 0 1 0 1 ...
summary(SAheart)

X = SAheart[,-c(1,6,11)]
pairs(X,pch=21,cex=0.8,
      bg=c("green","red")[unclass(factor(SAheart$chd))])
```

Algunas preguntas sobre estos datos

- El vector de observaciones para cada individuo, ¿tiene distribución normal multivariante?
- Podemos establecer alguna relación entre la tensión sanguínea y otras variables, como la obesidad, la edad, el consumo de tabaco y alcohol?
- Dado un nuevo individuo para el que se conocen todas las variables explicativas, ¿existe algún método para clasificarlo como paciente coronario o sano?
- Dado un nuevo individuo para el que se conocen todas las variables explicativas, ¿podemos estimar la probabilidad con la que sufra un infarto?

Referencias

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second edition. Springer. `https://web.stanford.edu/~hastie/ElemStatLearn/`