Práctica 3: Estadística descriptiva bivariante

Estadística

Grado en Inteligencia Artificial (UDC)

Índice

3. Estadística descriptiva bivariante

3.1. Variable estadística bidimensional

Como ya se comentó, cuando en el conjunto de datos contiene observaciones de múltiples variables (normalmente características de los individuos de una muestra), el análisis descriptivo suele comenzar con el análisis de estas variables de forma individual, sin embargo, normalmente resulta de mayor interés analizar estas variables de forma conjunta. Como las observaciones son características correspondientes a un mismo individuo, es de esperar que haya relación entre ellas, por lo que uno de los principales objetivos suele ser estudiar la posible relación (asociación) entre variables.

En esta práctica nos centraremos en el análisis simultáneo de dos variables estadísticas X e Y, por tanto, supondremos que disponemos de una muestra $\{(x_1, y_1), \dots, (x_n, y_n)\}$.

Individuo	X	Y
1	x_1	y_1
2	x_2	y_2
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

El par (X, Y) se denomina variable estadística bidimensional. En muchas ocasiones una de las variables es de mayor interés, en esos casos la **variable de interés** (o respuesta) se suele denotar por Y, y la otra variable X se suele denominar **variable explicativa** (o factor, si es cualitativa). Además, en muchas de las herramientas de R se suelen emplear fórmulas, del tipo `y ~ x`, para especificar esta posible relación.

Las variables X e Y pueden ser cualitativas (nominales o ordinales) o cuantitativas (discretas o continuas), dependiendo de la combinación de tipos de ambas variables tienen sentido unos análisis u otros. Al menos como punto de partida, normalmente se emplean alguno de los siguientes métodos descriptivos:

- Si una de las variables es cuantitativa (continua o discreta con muchos valores distintos) y la otra es categórica, se analizan las **distribuciones condicionadas** de la variable numérica para los distintos valores de la variable cualitativa, empleando:

- Gráficos de cajas.
- Medidas descriptivas.

- Si ambas variables son cualitativas (o cuantitativas discretas con pocos valores distintos), se suelen emplear:
  - Gráficos de barras.
  - **Tablas de contingencia** y medidas de asociación.
- Si ambas variables son cuantitativas (continuas o discreta con muchos valores distintos), se suelen emplear:
  - Gráficos de dispersión.
  - **Análisis de regresión** y correlación.

3.2. Análisis de una variable numérica y una categórica

En ocasiones interesar estudiar la distribución de una variable condicionada a que la otra tome un determinado valor (o valores). En esta sección supondremos que (X, Y) es una variable estadística bidimensional, tal que:

- Y es una variable cuantitativa.
- X es una variable cualitativa (o discreta) con k modalidades $c_1, c_2, \dots, c_k$.

Si sólo consideramos los individuos con $X = c_i$, hablaremos de la *distribución de la variable Y condicionada a $X = c_i$*. La variable condicionada se denotará por $Y|X = c_i$ (esta definición es válida cualesquiera que sean los tipos de las variables).

Se pueden emplear los métodos descritos en el tema anterior para estudiar las distribuciones condicionadas (en este caso podemos realizar un análisis univariante de la variable cuantitativa para cada modalidad de la variable cualitativa).

3.2.1. Gráficos de caja

Para analizar y comparar las distribuciones condicionadas son especialmente recomendables los gráficos de cajas. Podemos generar un gráfico de cajas empleando el método para `formula` (en lugar del método por defecto para vectores) de la función genérica `boxplot()`: `boxplot(formula, data, ...)`.

Por ejemplo, continuando con los datos de móviles de la práctica anterior, supongamos que nos interesa estudiar el `precio` dependiendo de la `marca` del teléfono:

```
load("movil.RData")
# str(movil)
# as.data.frame(attr(movil, "variable.labels")) # Etiquetas variables
boxplot(precio ~ marca, data = movil,
        ylab = "Precio (euros)", xlab = "Fabricante")
```

De estos gráficos se puede deducir mucha información:

1. Comparando las medianas podemos estudiar si el factor (la variable cualitativa) influye en la posición central de la respuesta (la variable cuantitativa).
   En este caso aparentemente hay grandes diferencias en el precio medio entre fabricantes, especialmente entre *Apple* y el resto (abusando de la notación podríamos escribir que mediana(precio|Apple) > mediana(precio|Xiaomi) > mediana(precio|Huawei)).
2. Comparando las alturas de las cajas podemos estudiar si hay diferencias en la variabilidad. En este caso, aparentemente la variabilidad del precio en *Huawei* es mayor y no hay grandes diferencias en

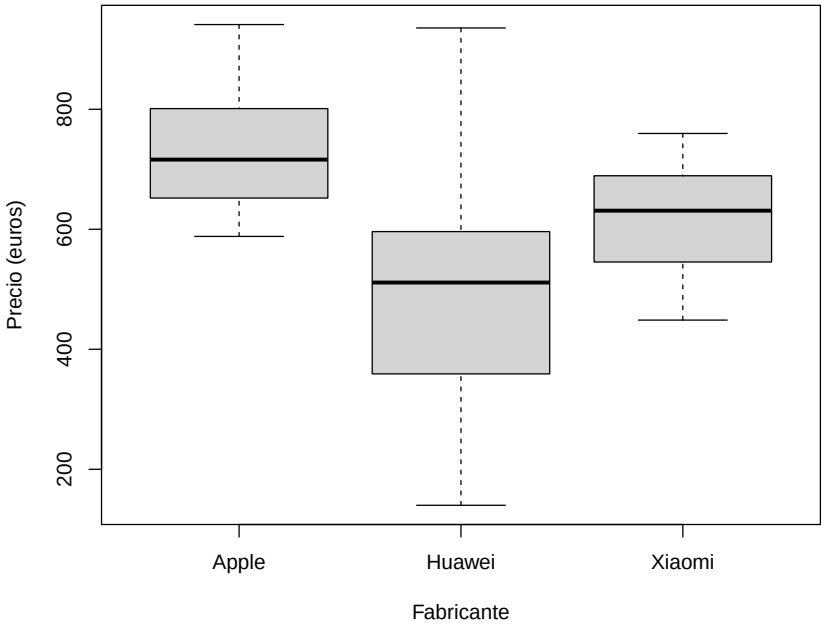


Figura 1: Gráfico de cajas (boxplot) de precio según fabricante.

la variabilidad entre los otros fabricantes (varibilidad(precio|Huawei > varibilidad(precio|Apple) = varibilidad(precio|Xiaomi)).

3. Analizando la forma de las cajas podemos estudiar la asimetría de las distribuciones condicionadas. En este caso son bastante simétricas (quizás en precio|Apple se observa una ligera asimetría positiva y en precio|Xiaomi una ligera asimetría negativa).
4. Finalmente podemos detectar la posible presencia de datos atípicos, que en este caso no se observan.

3.2.2. Medidas descriptivas

Podemos completar la información gráfica mediante estadísticos descriptivos. Para calcular medidas descriptivas de las distribuciones condicionadas podemos emplear la función `tapply(X, INDEX, FUN, ...)`. Por ejemplo, para obtener el precio medio por fabricante:

```
tapply(movil$precio, movil$marca, mean)
```

```
##    Apple   Huawei   Xiaomi
## 732.6312 497.3496 612.9751
```

También se podría emplear[1] las funciones `by(data, INDICES, FUN, ...)` o `aggregate(formula, data, FUN, ...)` (que ademas permiten operar sobre data.frames). Por ejemplo:

```
stat <- function(x) c(media = mean(x), mediana = median(x), sd = sd(x))
aggregate(precio ~ marca, movil, stat)
```

```
##    marca precio.media precio.mediana precio.sd
## 1  Apple    732.63122      716.08398 112.96108
```

[1] Además de que la función `boxplot()` devuelve de forma invisible los valores empleados en la construcción del gráfico.

```
## 2 Huawei    497.34956      511.22435 190.29315
## 3 Xiaomi    612.97508      630.99246  91.07622
```

```
# by(movil$precio, movil$marca, summary)
```

3.2.3. Ejercicio

Empleando los datos (modificados) de vehículos eléctricos `ecars` de la práctica anterior almacenados en el fichero *ecars2.RData*:

a) Estudiar si la carga rápida (`cargarapida`) influye en la distancia recorrida con carga completa (`dismax`).

b) Analizar la velocidad máxima (`velmax`) dependiendo del tipo de tracción (`traccion`).

c) Estudiar la distribución del precio (`logprecio`) dependiendo del tipo de vehículo (`carroceria2`).

3.3. Análisis de dos variables categóricas

Se pueden extender las herramientas para el análisis de una variable cualitativa al caso bivariante. En este caso tendríamos tablas de frecuencias bidimensionales, denominadas **tablas de contingencia**, que también podríamos obtener con la función `table()` y podríamos representar mediante gráficos de barras con la función `barplot()` (como ya se comentó, normalmente comenzaríamos por analizar los gráficos).

3.3.1. Tablas de contingencia

Sea (X, Y) una variable estadística bidimensional, tal que:

- X es una variable cualitativa (o discreta) con k modalidades $c_1, c_2, \dots, c_k$.
- Y es una variable cualitativa (o discreta) con l modalidades $c'_1, c'_2, \dots, c'_l$.

Toda la información proporcionada por los sujetos de la muestra se puede resumir en una **Tabla de contingencia**, una tabla de frecuencias bidimensional en la que las filas se corresponden con una de las variable categóricas y las columnas con la otra (la recomendación sería poner en columnas la variable con menor número de modalidades). Para cada combinación de filas y columnas se calcula la frecuencia de la correspondiente combinación de modalidades (cada par de posibles valores (c_i, c'_j) puede repetirse una o más veces en la muestra). Al igual que en el caso univariante podemos considerar frecuencias absolutas o relativas (estás últimas también se suelen mostrar en escala de porcentajes):

- **Frecuencia absoluta** del par (c_i, c'_j) n_{ij} = número de individuos con $X = c_i$ e $Y = c'_j$.
- **Frecuencia relativa** del par $(c_i, c'_j) = f_{ij} = \dfrac{n_{ij}}{n}$ = proporción de individuos con $X = c_i$ e $Y = c'_j$.

A partir de ellas podemos estudiar la **distribución conjunta** de ambas variables. Por ejemplo, la tabla de contingencia de frecuencias absolutas sería de la forma:

X \ Y	c'_1	c'_2	...	c'_j	...	c'_l
c_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1l}
c_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2l}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$
c_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{il}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$
c_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{kl}

Como ya se comentó, podemos obtener la tabla de contingencia de frecuencias absolutas con el comando `table(x, y)`. Por ejemplo:

```
frec <- table(movil$nsatisfa, movil$marca)
frec
```

```
##
##         Apple Huawei Xiaomi
##   Bajo      2      7      5
##   Medio     5      4      7
##   Alto      5      7      8
```

A partir de la que podemos calcular las frecuencias relativas (proporciones) o los porcentajes de las categorías, por ejemplo empleando la función `prop.table()`:

```
# Frecuencias relativas
prop.table(frec)   # frec/sum(frec)
```

```
##
##         Apple Huawei Xiaomi
##   Bajo   0.04   0.14   0.10
##   Medio  0.10   0.08   0.14
##   Alto   0.10   0.14   0.16
```

```
# Porcentajes
porc <- 100*prop.table(frec)
porc
```

```
##
##         Apple Huawei Xiaomi
##   Bajo      4     14     10
##   Medio    10      8     14
##   Alto     10     14     16
```

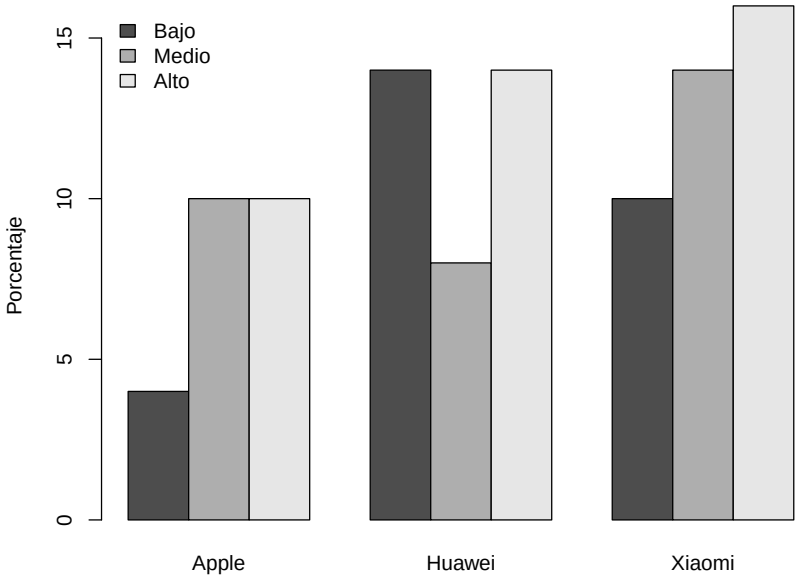
Es decir, hay 8 individuos con móviles de la marca Xiaomi y un nivel de satisfacción alto (la combinación de categorías más frecuente, con un 16 % de las observaciones). La combinación de categorías menos frecuente es un nivel de satisfacción bajo con un móvil Apple con un 4 % de los individuos (2 casos).

Como también se comentó, podemos representar la distribución conjunta mediante un gráfico de barras con la función `barplot()` (lo que en general permitiría detectar con mayor comodidad las combinaciones de categorías más frecuentes y las menos frecuentes). En este caso nos pueden interesar añadir los argumentos:

- `beside = TRUE`: para generar un gráfico de barras agrupado (si el número de categorías es pequeño). Por defecto representa barras apiladas (`beside = FALSE`).
- `legend.text = TRUE` (o un vector de etiquetas de valores): para añadir una leyenda (se pueden incluir argumentos adicionales con el parámetro `args.legend`).

Por ejemplo:

```r
barplot(porc, ylab = "Porcentaje", beside = TRUE, legend.text = TRUE,
        args.legend = list(x = "topleft", bty = "n"))
```



3.3.2. Ejercicio

Continuando con los datos (modificados) de vehículos eléctricos `ecars` del ejercicio anterior:

a) Obtener la tabla de contingencia y el gráfico de barras del tipo de tracción (`traccion`) por carga rápida (`cargarapida`).

b) Realizar un análisis descriptivo de la distribución conjunta de carga rápida (`cargarapida`) y segmento de mercado (`segmento`).

3.3.3. Distribuciones marginales

En muchas ocasiones se suelen incluir en la tabla de contingencia los totales por filas y columnas, denominadas **distribuciones marginales**:

$X \quad Y$	c'_1	c'_2	...	c'_j	...	c'_l	$Total$
c_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1l}	$n_{1\cdot}$
c_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2l}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
c_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{il}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
c_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{kl}	$n_{k\cdot}$
$Total$	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot j}$	...	$n_{\cdot l}$	n

Donde:

- $n_{i\cdot} = n_{i1} + n_{i2} + ... + n_{il}$ = número de individuos con $X = c_i$.

- $n_{\cdot j} = n_{1j} + n_{2j} + ... + n_{kj}$ = número de individuos con $Y = c'_j$.
- n = número (total) de individuos.

El caso de frecuencias relativas sería análogo:

- $f_{i\cdot} = \frac{n_{i\cdot}}{n} = f_{i1} + f_{i2} + ... + f_{il}$ = proporción de individuos con $X = c_i$.
- $f_{\cdot j} = \frac{n_{\cdot j}}{n} = f_{1j} + f_{2j} + ... + f_{kj}$ = proporción de individuos con $Y = c'_j$.
- El total sería 1 en lugar de n.

Las distribuciones marginales son simplemente las distribuciones de frecuencias unidimensionales de las variables X e Y (su nombre se debe a que se obtienen añadiendo en los márgenes de la tabla las sumas de las frecuencias de la distribución conjunta) y son de utilidad para estudiar las variables de forma individual. Por tanto, a partir de la tabla de contingencia podemos realizar los análisis descriptivos univariantes descritos en la práctica anterior.

Podemos obtener estas distribuciones con la función `addmargins()`. Por ejemplo:

```
addmargins(frec)
```

```
##
##         Apple Huawei Xiaomi Sum
##   Bajo      2      7      5  14
##   Medio     5      4      7  16
##   Alto      5      7      8  20
##   Sum      12     18     20  50
```

```
addmargins(porc, margin = 2, FUN = list(Total = sum))
```

```
##
##         Apple Huawei Xiaomi Total
##   Bajo      4     14     10    28
##   Medio    10      8     14    32
##   Alto     10     14     16    40
```

3.3.4. Distribuciones condicionadas

Para estudiar si hay relación (asociación) entre las variables nos interesará comparar las distribuciones condicionadas (que se definen como en la sección anterior, aunque en este caso son cualitativas). Si sólo consideramos los individuos con $Y = c'_j$, hablaremos de la distribución de la variable X condicionada a $Y = c'_j$ (que denotaremos por $X|Y = c'_j$).

La tabla de frecuencias absolutas de X condicionada a $Y = c'_j$, será:

$X\|Y = c'_j$	c_1	$\cdots$	c_i	$\cdots$	c_k
	$n_{1/j}$	$\cdots$	$n_{i/j}$	$\cdots$	$n_{k/j}$

donde $n_{i/j} = n_{ij}$. Es decir, es simplemente la columna j de la tabla de contingencia de frecuencias absolutas. Análogamente obtendríamos la tabla de frecuencias absolutas de $Y|X = c_i$.

Por ejemplo:

```
# Tabla de frecuencias absolutas del nivel de satisfacción condicionada a fabricante Apple
frec[, "Apple"]
```

```
##  Bajo Medio  Alto
##     2     5     5
```

```r
# Tabla de frecuencias absolutas de fabricante condicionada a nivel de satisfacción alto
frec["Alto", ]
```

```
##  Apple Huawei Xiaomi
##      5      7      8
```

A partir de estos valores podríamos obtener las frecuencias relativas. Por ejemplo la tabla de frecuencias relativas de X condicionada a $Y = c'_j$, será de la forma:

$X\|Y = c'_j$	c_1	$\cdots$	c_i	$\cdots$	c_k
	$f_{1/j}$	$\cdots$	$f_{i/j}$	$\cdots$	$f_{k/j}$

donde $f_{i/j} = \frac{n_{ij}}{n_{\cdot j}} = \frac{f_{ij}}{f_{\cdot j}}$. De forma análoga se obtendrían las frecuencias relativas de cada modalidad de Y condicionadas a una modalidad de X.

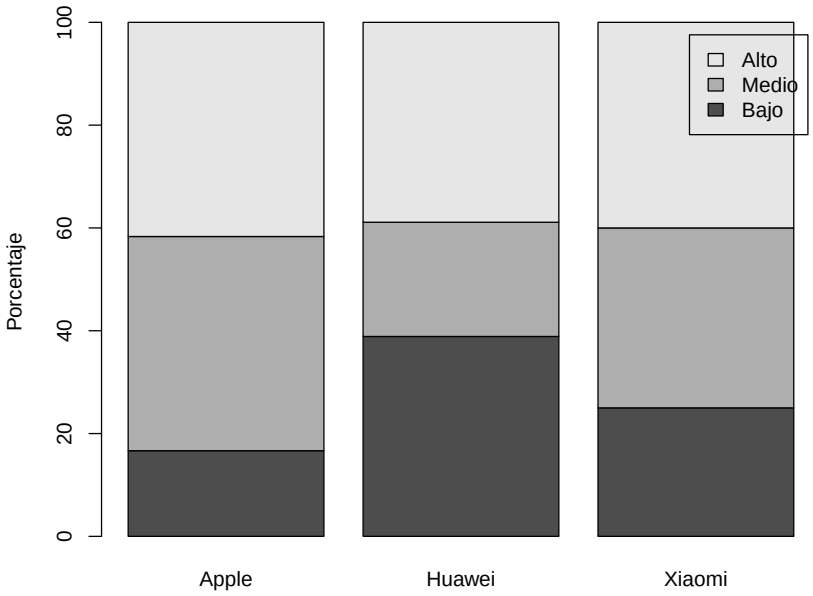
Para calcular de forma simultánea las tablas de frecuencias relativas de las distribuciones condicionadas podemos emplear el argumento `margin` (índice sobre el que se condiciona) de la función `prop.table()`.

Por ejemplo, podemos obtener y representar las distribuciones del nivel de satisfacción condicionadas a los distintos fabricantes:

```r
# Distribución del nivel de satisfacción según fabricante
porc.cond <- 100*prop.table(frec, 2)
porc.cond
```

```
##
##            Apple   Huawei   Xiaomi
##   Bajo  16.66667 38.88889 25.00000
##   Medio 41.66667 22.22222 35.00000
##   Alto  41.66667 38.88889 40.00000
```

```r
# Gráfico de barras apilado
barplot(porc.cond, ylab = "Porcentaje", legend.text = TRUE)
```




```
# La suma por columnas es el 100%
# addmargins(porc.cond, 1)
```

Si no hay relación (asociación) entre las variables las distribuciones condicionales deberían ser similares a la correspondiente distribución marginal. A efectos ilustrativos podemos repetir el ejemplo anterior pero añadiendo la distribución marginal del nivel de satisfacción:

```
frec2 <- addmargins(frec, 2)
# frec2 <- addmargins(frec, margin = 2, FUN = list(Total = sum))
100*prop.table(frec2, 2)
```

```
##
##            Apple   Huawei   Xiaomi      Sum
##   Bajo  16.66667 38.88889 25.00000 28.00000
##   Medio 41.66667 22.22222 35.00000 32.00000
##   Alto  41.66667 38.88889 40.00000 40.00000
```

3.3.5. Medidas de asociación

Para cuantificar el grado de asociación nos interesaría calcular medidas descriptivas que, por ejemplo, tomen el valor 0 cuando el nivel de asociación sea nulo y el valor 1 cuando el nivel de asociación sea máximo.

Si no hay ninguna asociación entre las variables las tablas de frecuencias relativas de $Y|X = c_1, \dots, Y|X = c_k$ serán iguales (a las frecuencias relativas marginales de Y). Es decir:

$$f_{i/j} = \frac{f_{ij}}{f_{\cdot j}} = f_{i\cdot} \text{ para todo } i, j.$$

De donde se deduce que $f_{ij} = f_{i\cdot} f_{\cdot j}$, o equivalentemente:

$$n_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n}$$

para todo i, j. Los valores:

$$e_{ij} = \frac{n_{i.} n_{.j}}{n}$$

se denominan **frecuencias esperadas bajo independencia**.

Las medidas de asociación más simples (válidas para todo tipo de variables cualitativas) están basadas en el estadístico chi-cuadrado de Pearson (1900):

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

Este estadístico mide la distancia entre las frecuencias observadas y las esperadas bajo independencia, y toma el valor 0 cuando no hay asociación entre las variables (estudiaremos con mayor profundidad este estadístico en la parte de inferencia estadística).

Podemos obtener medidas de asociación reescalando esta distancia de forma que tome valores entre 0 y 1. Las más conocidas son:

- El coeficiente de contingencia:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

  (su valor máximo puede ser menor de 1).

- La V de Cramer:

$$V = \sqrt{\frac{\chi^2}{n \cdot (m-1)}}$$

  donde m es el mínimo del número de filas y de columnas (puede tomar el valor 1, asociación completa, en tablas de cualquier dimensión).

Podemos calcular estas medidas a partir de los resultados de la función `chisq.test()` del paquete base de R, pero puede resultar más cómodo emplear la función `assocstats()` del paquete `vcd`:

```
library(vcd)
```

```
## Loading required package: grid
```

```
assocstats(frec)
```

```
##                     X^2 df P(> X^2)
## Likelihood Ratio 2.3815  4  0.66597
## Pearson          2.3353  4  0.67435
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.211
## Cramer's V        : 0.153
```

```
# Empleando chisq.test:
chisq.stat <- chisq.test(frec)$statistic
```

```
## Warning in chisq.test(frec): Chi-squared approximation may be incorrect
```

```
chisq.stat # Estadístico chi-cuadrado
```

```
## X-squared
##  2.335317
```

```
names(chisq.stat) <- NULL
# Coeficiente de contingencia
sqrt(chisq.stat / (chisq.stat + sum(frec)))
```

```
## [1] 0.2112397
```

```
# V de Cramer
sqrt(chisq.stat / (sum(frec) * (min(dim(frec)) - 1)))
```

```
## [1] 0.1528175
```

En este caso el grado de asociación entre satisfacción y fabricante es bastante bajo.

En el caso de variables ordinales (o discretas) podemos emplear medidas de asociación que aprovechen la información adicional que proporciona la ordenación. En la siguiente sección se mostrarán como ejemplo el coeficiente de correlación de Spearman y la tau-b de Kendall (que aparecen como alternativa al coeficiente de correlación lineal de Pearson).

3.3.6. Ejercicio

Continuando con los datos (modificados) de vehículos eléctricos `ecars` de ejercicios anteriores:

a) Estudiar la distribución de carga rápida (`cargarapida`) condicionada al tipo de tracción (`traccion`).

b) Realizar un análisis descriptivo completo (incluyendo medidas de asociación) de tipo de tracción (`traccion`) y tipo de vehículo (agrupado) (`carroceria2`).

3.4. Análisis de dos variables numéricas

En esta sección nos centraremos en el caso de dos variables numéricas, y supondremos que Y es la variable de interés o respuesta (también denominada variable dependiente en este contexto) y X es la variable explicativa (también denominada predictor, variable regresora o variable independiente). Nos interesa estudiar la posible relación entre estas dos variables (si la distribución de Y depende de X y en caso afirmativo, cuál es la relación funcional entre ellas).

Es importante tener en cuenta que relación no implica causalidad, es lo que se conoce como relación espuria (o correlación espuria). La aparente relación puede ser debida a la casualidad o a otras variables que no se tienen en cuenta (denominadas *factores de confusión* o *variables ocultas*)[2].

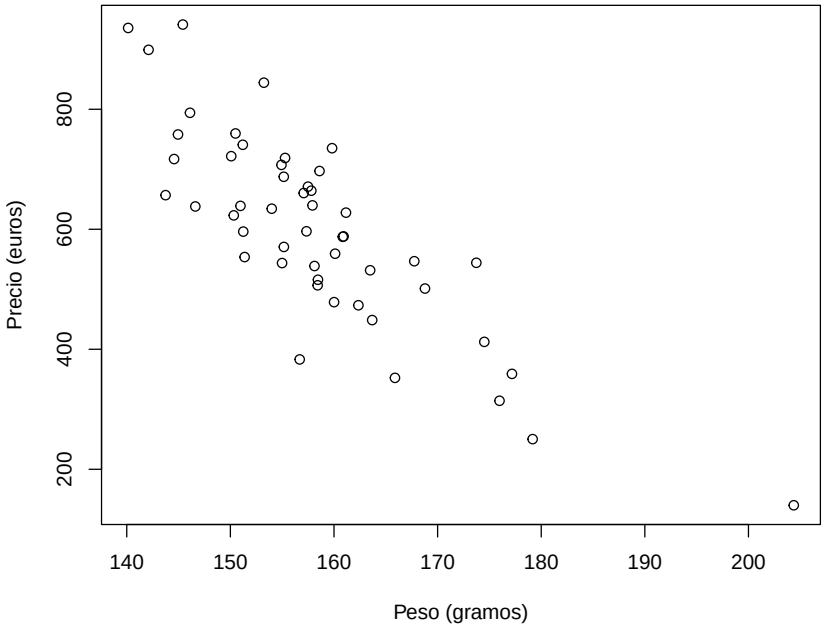
3.4.1. Gráfico de dispersión

Como primer paso la recomendación es generar un gráfico de dispersión de Y sobre X, en el que se representa cada par de observaciones (x_i, y_i) como un punto en el plano cartesiano. Si se observa alguna forma concreta en la nube de puntos, indicaría que hay algún tipo de relación entre las variables. Si los puntos de la nube se agrupan en torno a una recta, diremos que las variables parecen estar relacionados linealmente (la relación es lineal, aparentemente).

En R podemos generar este gráfico con el comando `plot(x, y)`. Por ejemplo:

```
plot(movil$peso, movil$precio, xlab = "Peso (gramos)", ylab = "Precio (euros)")
```

[2] Ver por ejemplo el blog spurious-correlations o Chocolate creates Nobel prize winners.



En este caso aparentemente hay una relación lineal negativa, al aumentar el peso disminuye el precio (de forma lineal).

3.4.2. Medidas descriptivas

Podemos emplear los estadísticos descriptivos univariantes descritos en la práctica anterior para analizar las variables por separado.

Por ejemplo, podemos calcular las medias $\overline{x}$ e $\overline{y}$ de ambas variables. El vector:

$$\begin{pmatrix} \overline{x} \\ \overline{y} \end{pmatrix}$$

se denomina **vector de medias** de la variable (X, Y).

Análogamente podemos calcular las varianzas s_X^2 y s_Y^2. En este caso además nos interesará una medida de la variabilidad conjunta de ambas variables, la **covarianza** de (X, Y):

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}$$

Podemos obtener la covarianza con la función `cov()`. Por ejemplo:

```
cov(movil$peso, movil$precio)
```

```
## [1] -1533.222
```

Se denomina matriz de varianzas-covarianzas de la variable (X, Y) a la matriz:

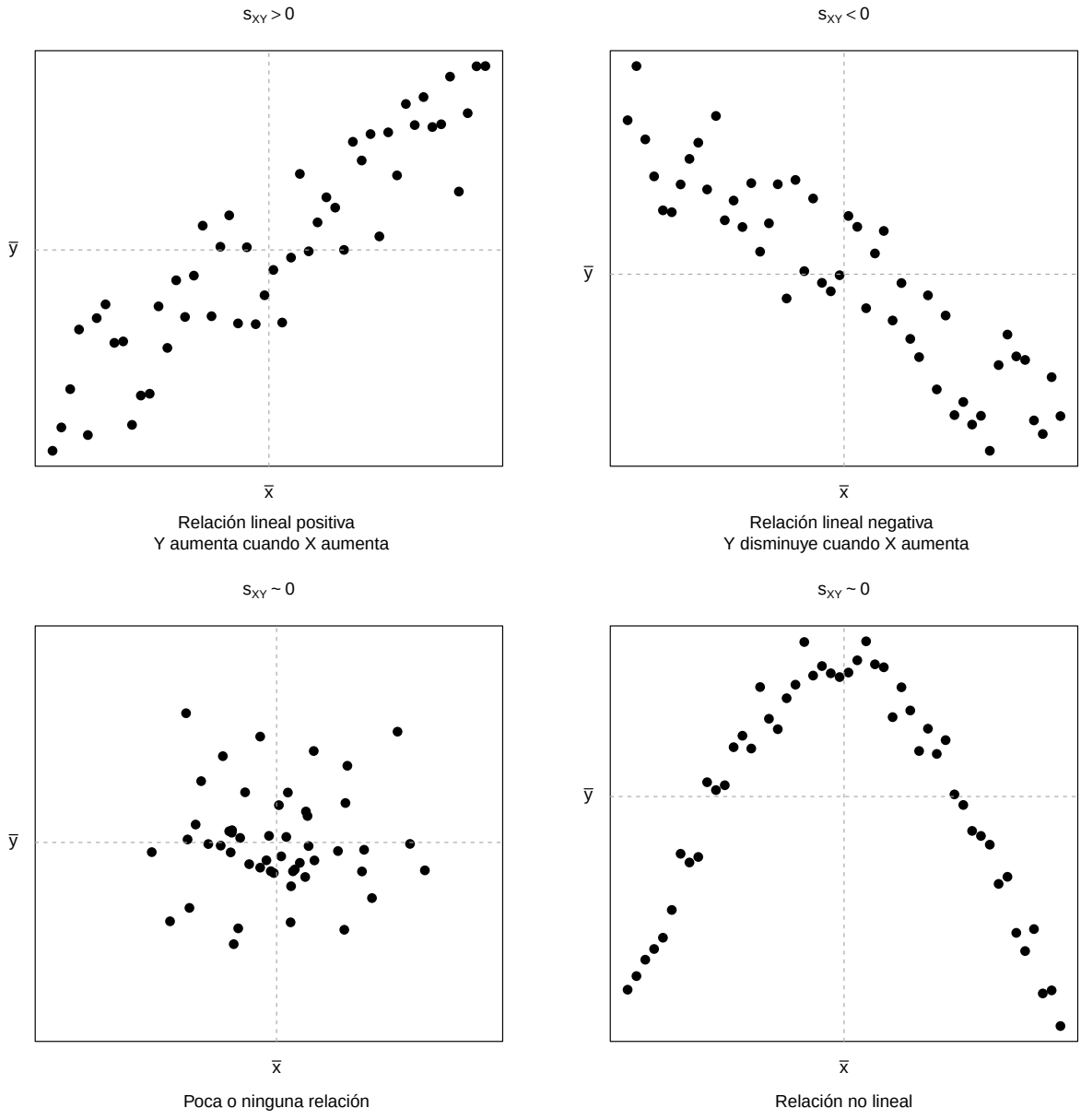
$$S = \begin{pmatrix} s_X^2 & s_{XY} \\ s_{XY} & s_Y^2 \end{pmatrix}$$

La función `cov()` devuelve esta matriz cuando el primer argumento es un data.frame o una matriz:

```
cov(movil[c("peso", "precio")])
```

```
##              peso    precio
## peso     126.6185 -1533.222
## precio -1533.2224 26891.581
```

La covarianza es una medida del grado de relación lineal:



Sin embargo, el valor de la covarianza depende de la escala de las variables (por lo que resulta complicado saber cuando es grande o próxima a cero). Para medir el grado de dependencia o relación (lineal) entre las variables es preferible reescalar este valor (por ejemplo de forma que su valor máximo sea conocido).

3.4.3. El coeficiente de correlación lineal

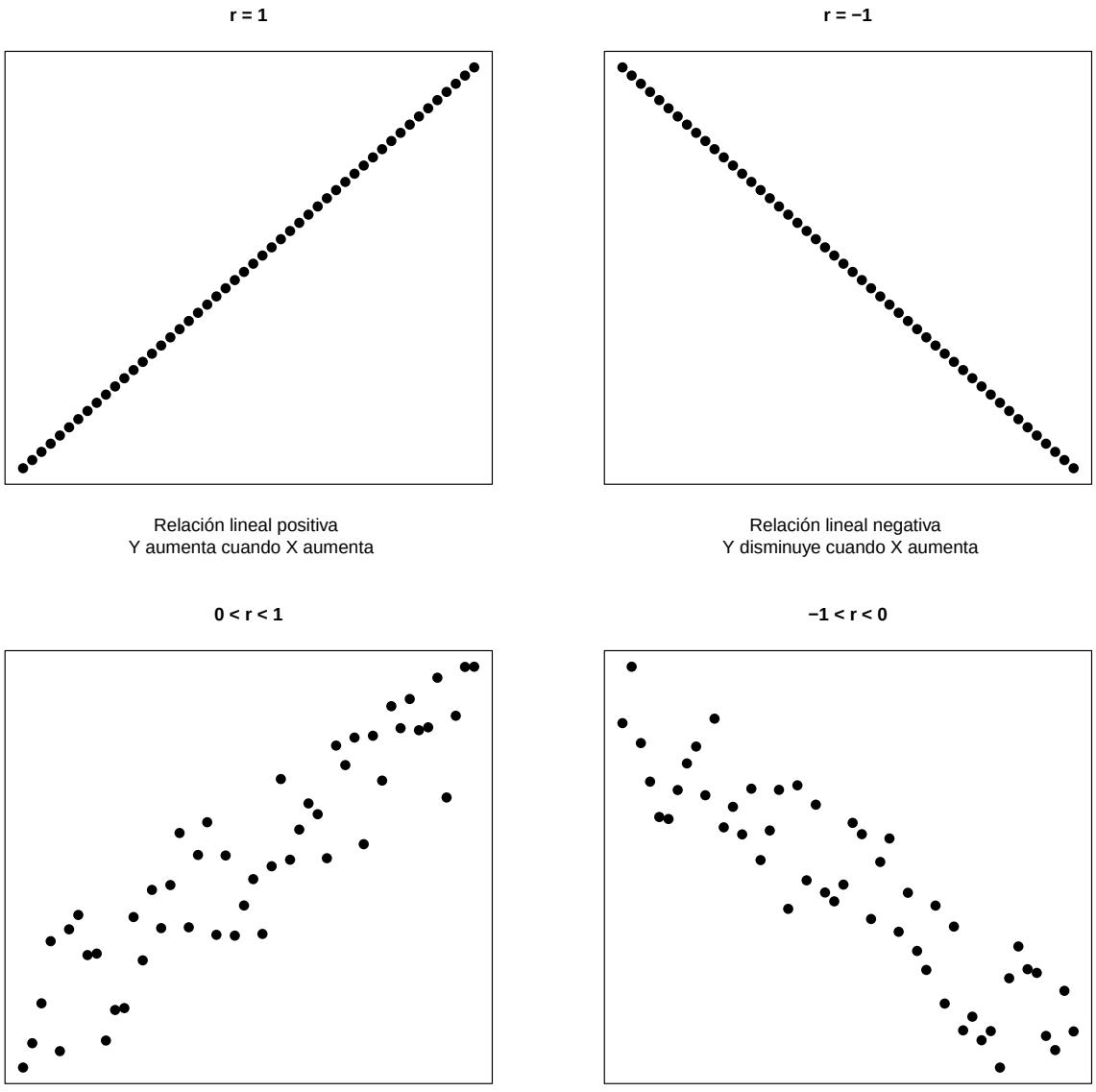
El coeficiente de correlación lineal de Pearson:

$$r = \frac{s_{XY}}{s_X s_Y}$$

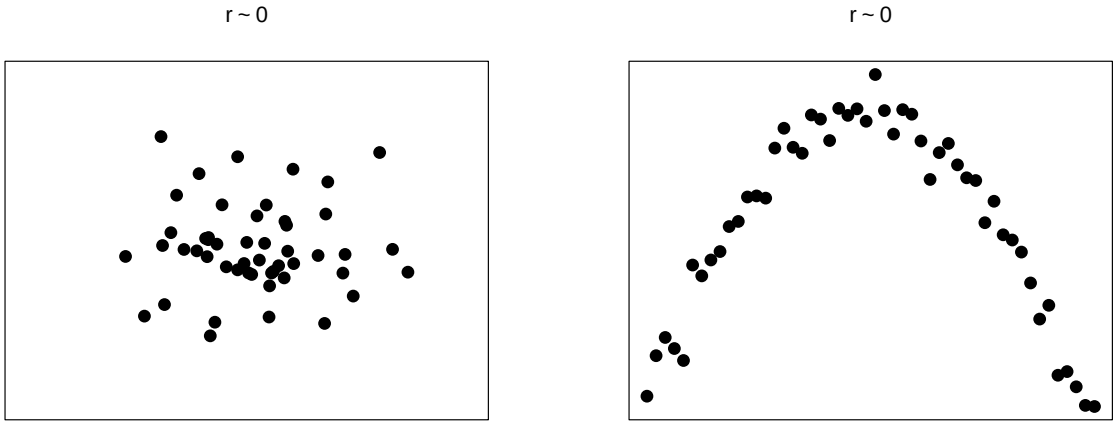
es una medida adimensional de la **relación lineal** entre dos variables cuantitativas (no depende de las unidades de medida). Siempre toma valores entre -1 y 1:

$$-1 \leq r \leq 1$$

Si $r = \pm 1$ hay una relación lineal exacta entre las dos variables:



Si $r = 0$ no hay relación lineal entre las variables (puede haber una relación no lineal) y se dice que las variables son **incorreladas**:



Podemos obtener el coeficiente de correlación con la función `cor()`. Por ejemplo:

```
# Coeficiente de correlación
cor(movil$peso, movil$precio)
```

```
## [1] -0.8308993
```

```
# Matriz de correlaciones
cor(movil[c("peso", "precio")])
```

```
##              peso     precio
## peso    1.0000000 -0.8308993
## precio -0.8308993  1.0000000
```

Además, el coeficiente de correlación lineal:

- No se ve afectado por transformaciones lineales.
- No es una medida robusta, puede verse seriamente afectado por observaciones atípicas.

También hay modificaciones de este coeficiente más adecuadas para el caso de distribuciones asimétricas, o si hay observaciones atípicas, o si alguna de las variables es discreta o incluso solo ordinal, como el coeficiente de correlación por rangos de Spearman (transforma las variables a rangos) o la tau-b de Kendall (mide la concordancia entre pares de observaciones). También podemos obtener estas medidas la función `cor()`, estableciendo el parámetro `method = "spearman"` o `method = "kendall"`, respectivamente.

3.4.4. Ejercicio

Continuando con los datos (modificados) de vehículos eléctricos `ecars` de ejercicios anteriores:

a) Generar el gráfico de dispersión de la eficiencia (`eficiencia`) sobre la velocidad de carga (`velcarga`). Calcular la covarianza y la correlación entre ambas variables (Nota: si hay datos faltantes se puede emplear `use = çomplete.obs"` al llamar a `cov()` o `cor()` para no tenerlas en cuenta en los cálculos).

b) Estudiar si el precio (`logprecio`) es de utilidad para explicar la velocidad máxima (`velmax`).

3.4.5. Regresión y correlación

El análisis de **regresión** (1889, Francis Galton, Natural inheritance) es una técnica estadística centrada en el estudio de las posibles relaciones entre variables con el propósito de establecer una relación funcional entre

ellas (un modelo matemático que relacione una respuesta con un conjunto de variables explicativas). Por **correlación** se entiende el grado de asociación entre dos (o más) variables. En este apartado nos centraremos únicamente en regresión simple (dos variables numéricas), desde un punto de vista descriptivo (se volverá a tratar más adelante en la parte de inferencia estadística).

Nos interesará especialmente estudiar si hay una relación lineal entre las dos variables (la relación más simple). Si suponemos que la variable respuesta Y y la variable explicativa X están relacionadas linealmente podemos obtener la *recta de regresión mínimo cuadrática* (aproximación de la relación entre ambas variables) y estudiar la *bondad del ajuste* (tratar de medir lo adecuada que es la recta ajustada para explicar la respuesta; en este caso será una medida del grado de relación lineal).

En general tendríamos tres posibles situaciones:

- **Relación exacta** (o **funcional**): la variable explicativa determina totalmente el valor de la respuesta:

$$Y = m(X)$$

- **Independencia:** la variable explicativa no aporta ninguna información sobre la respuesta.
- **Relación estadística** o **estocástica**: la variable explicativa permiten predecir en mayor o menor grado el valor de la respuesta:

$$Y = m(X) + \varepsilon$$

Se puede explicar la respuesta mediante una función de la **variable explicativa**, más un **término de error** ε (que recogería el efecto conjunto de otras variables no consideradas).

Consideraremos el caso más simple, lo que se conoce como **regresión lineal simple**. Supondremos que la variable respuesta Y y la variable explicativa X están relacionadas linealmente:

$$Y = a + bX + \varepsilon$$

El **objetivo principal** es, a partir de los valores observados:

$$\{(x_i, y_i) : i = 1, \cdots, n\},$$

$$y_i = a + bx_i + \varepsilon_i,$$

aproximar la recta de regresión:

$$y = a + bx$$

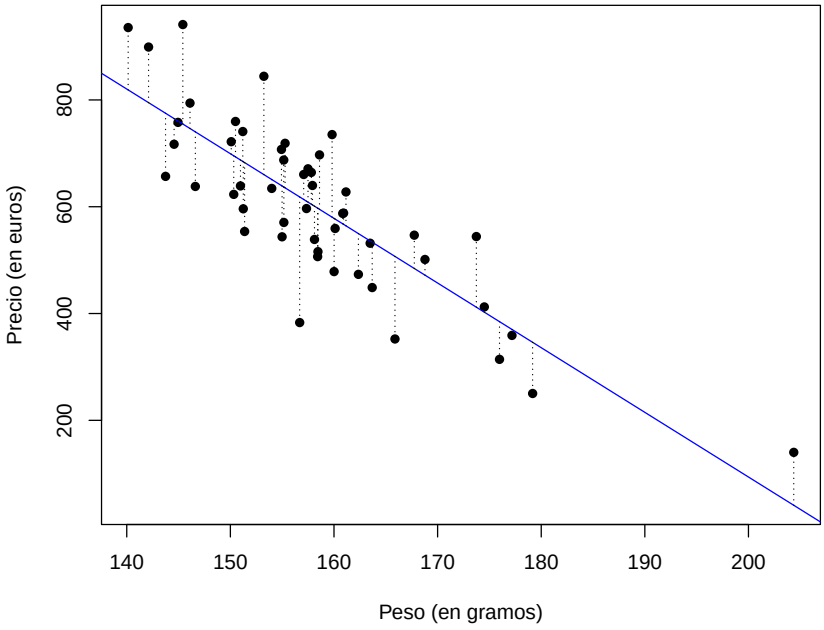
(es decir, aproximar los parámetros a y b).

Como criterio de ajuste (para buscar la recta con la que se predice mejor la respuesta) se empleará el método de mínimos cuadrados.

3.4.6. Recta de regresión mínimo cuadrática

Es la recta $\widehat{y} = \widehat{a} + \widehat{b}x$ que **minimiza la suma de los cuadrados de los errores**:

$$\min_{a,b} \sum_{i=1}^{n} (y_i - (a + bx_i))^2 = \min_{a,b} \sum_{i=1}^{n} e_i^2$$



Puede verse fácilmente[3] que:

$$\hat{a} = \overline{y} - \hat{b}\overline{x}$$
$$\hat{b} = \frac{s_{XY}}{s_X^2}$$

Entonces la ecuación de la recta de regresión mínimo cuadrática de Y sobre X puede expresarse como:

$$\hat{y} = \bar{y} + \frac{s_{XY}}{s_X^2}(x - \bar{x})$$

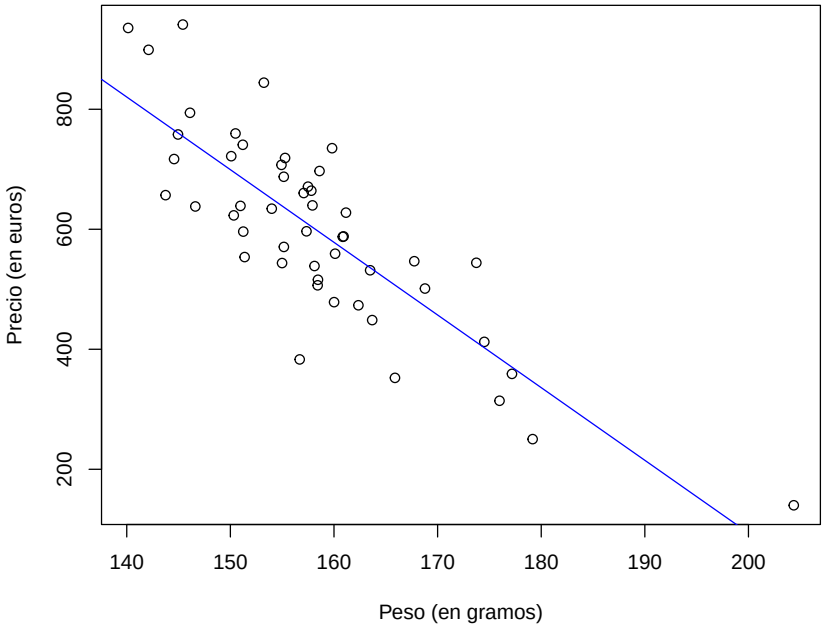
NOTAS:

- La recta de regresión mínimo cuadrática siempre pasa por el punto $(\bar{x}, \bar{y})$.
- La recta de regresión de Y sobre X $(Y|X)$ no coincide con la recta de regresión de X sobre Y $(X|Y)$, salvo relación lineal perfecta.

En R podemos realizar un ajuste de un modelo lineal con la función `lm()`. Por ejemplo:

```
fit <- lm(precio ~ peso, data = movil)
plot(movil$peso, movil$precio,
     xlab = "Peso (en gramos)", ylab = "Precio (en euros)")
abline(fit, col = "blue")
```

[3] Calculando las derivadas parciales de la suma de cuadrados, igualando a cero y resolviendo el sistema.




```
fit
```

```
##
## Call:
## lm(formula = precio ~ peso, data = movil)
##
## Coefficients:
## (Intercept)         peso
##     2515.65       -12.11
```

En este caso, la recta de regresión mínimo cuadrática para explicar el precio a partir del peso es:

$$\widehat{precio} = 2515{,}65 - 12{,}11 \cdot peso$$

.

Podemos interpretar los coeficientes:

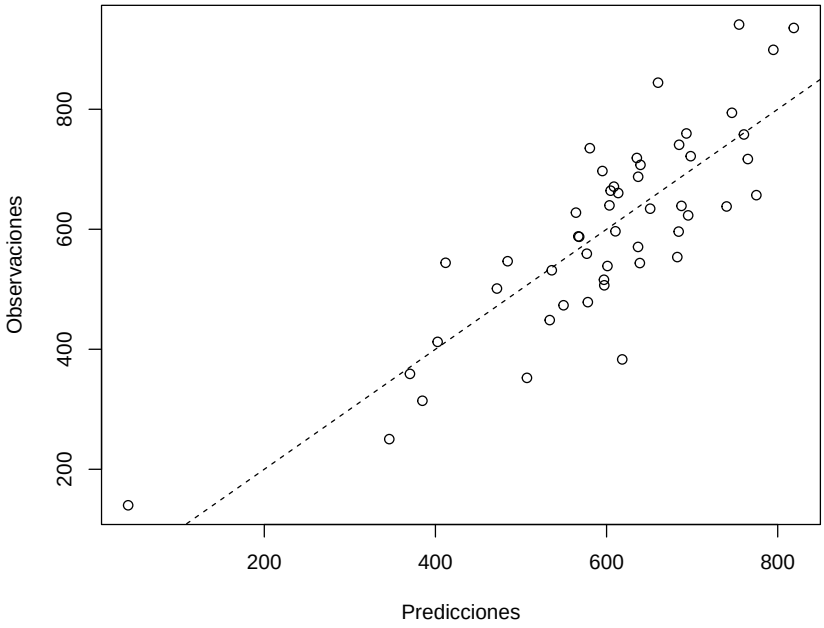
- $\hat{a}$: predicción de Y cuando $X = 0$ (no suele ser de gran interés).
- $\hat{b}$: Incremento en (la predicción de) Y cuando X aumenta una unidad.

En este caso, por cada incremento de un gramo de peso, la predicción del precio disminuye en 12.11 euros.

Sustituyendo en la recta ajustada la variable explicativa por un valor obtendríamos la correspondiente predicción de la respuesta. Podemos obtener estas predicciones con la función genérica `predict()`. Por defecto se obtienen las predicciones para los valores observados:

$$\hat{y}_i = \hat{a} + \hat{b}x_i$$

```
pred <- predict(fit)
# Gráfico de predicciones frente a observaciones
plot(pred, movil$precio, xlab = "Predicciones", ylab = "Observaciones")
abline(a = 0, b = 1, lty = 2) # recta x = y
```



Podemos obtener predicciones para nuevos valores empleando el parámetro `newdata` (que debe ser un data.frame; ver `?predict.lm`):

```
predict(fit, newdata = data.frame(peso = c(150, 200)))
```

```
##         1         2
## 699.30159  93.85195
```

3.4.7. Bondad del ajuste

Las diferencias entre valores observados y predicciones:

$$y_i - (\hat{a} + \hat{b}x_i) = y_i - \hat{y}_i = e_i$$

se denominan **residuos** (de media 0). Su varianza:

$$s_R^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

es una medida de la variabilidad de los datos respecto a la recta, denominada **varianza residual** (aunque también depende de la escala de la respuesta).

Una medida de la bondad del ajuste (evaluación global de la recta de regresión) es **el coeficiente de determinación**:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{s_E^2}{s_Y^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{s_R^2}{s_Y^2}$$

que se puede interpretar como la **proporción de variabilidad** (de la respuesta) **explicada por la regresión**.

Se verifica que $0 \leq R^2 \leq 1$:

- Si $R^2 = 1$ todas las observaciones están en la recta de regresión (lo explica todo)

- Si $R^2 = 0$ la recta de regresión no explica nada

Podemos obtener este coeficiente empleando la función `summary()`. Por ejemplo:

```
summary(fit)$r.squared
```

```
## [1] 0.6903937
```

La recta ajustada explicaría un 69 % de la variabilidad del precio, por lo tanto el ajuste es regular/bueno.

En el caso de regresión lineal simple, se puede interpretar del coeficiente de determinación a partir del coeficiente de correlación lineal de Pearson. Teniendo en cuenta que

$$\hat{y}_i = \bar{y} + \hat{b}(x_i - \bar{x}),$$

se puede expresar el coeficiente de determinación como:

$$R^2 = \hat{b}^2 \frac{s_X^2}{s_Y^2} = \frac{s_{XY}^2}{s_X^2 s_Y^2},$$

que resulta ser el cuadrado del coeficiente de correlación lineal de Pearson $r = \frac{s_{XY}}{s_X s_Y}$.

Por ejemplo:

```
cor(movil$peso, movil$precio)^2
```

```
## [1] 0.6903937
```

NOTA:

$$r = 0 \Leftrightarrow s_{XY} = 0 \Leftrightarrow \hat{b} = 0$$

3.4.8. Ejercicio

Continuando con los datos (modificados) de vehículos eléctricos `ecars` de ejercicios anteriores:

a) Obtener la recta de regresión mínimo cuadrática de velocidad máxima (`velmax`) sobre el precio en escala logarítmica (`logprecio`), representarla gráficamente y estudiar la bondad del ajuste. Emplearla para predecir la velocidad máxima de un coche eléctrico con un precio de 50000 euros.

b) Estudiar si la distancia máxima con carga completa (`dismax`) es de utilidad para explicar la velocidad de carga (`velcarga`). En caso afirmativo emplearla para predecir los datos faltantes de `velcarga`.

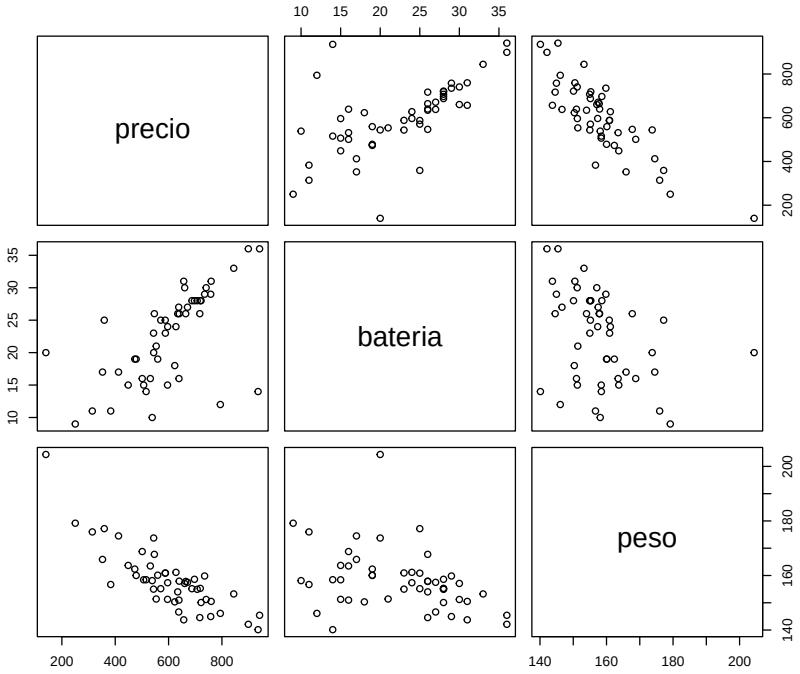
```
# load("ecars2.RData")
index <- which(is.na(ecars$velcarga))
ecars[index, ]
# Nota: son los 5 sin carga rápida (cuidado)
```

3.5. Estadística descriptiva multivariante

Los métodos descriptivos anteriores se pueden extender al caso multivariante (aunque se han desarrollado métodos específicos para analizar simultáneamente muchas variables). A continuación se muestran algunos ejemplos solo con fines ilustrativos.

Por ejemplo, en el caso de más de dos variables numéricas podemos emplear un gráfico de dispersión matricial:

```
plot(movil[5:7])
```



y calcular la matriz de correlaciones:

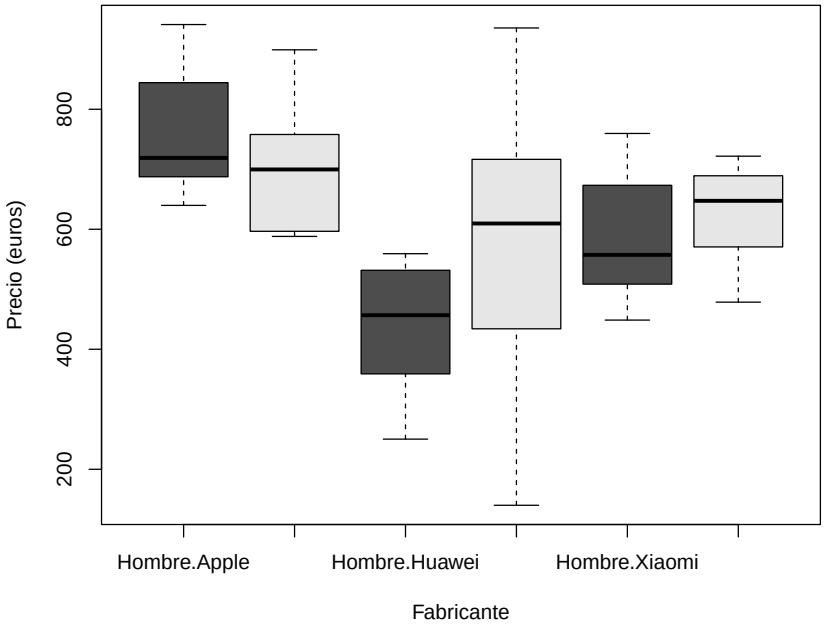
```
mcor <- cor(movil[c(5:7)])
print(mcor, digits = 2)
```

```
##         precio bateria  peso
## precio    1.00    0.61 -0.83
## bateria   0.61    1.00 -0.37
## peso     -0.83   -0.37  1.00
```

En el caso de una variable numérica (respuesta) y varias variables categóricas (factores), podemos generar gráficos de cajas agrupados[4]:

```
boxplot(precio ~ sexo + marca, data = movil, col = gray.colors(2),
        ylab = "Precio (euros)", xlab = "Fabricante")
```

[4]En el caso de tres factores se recomendaría generar uno por cada nivel del tercer factor. En los análisis gráficos no se suelen considerar más de tres factores simultáneamente.



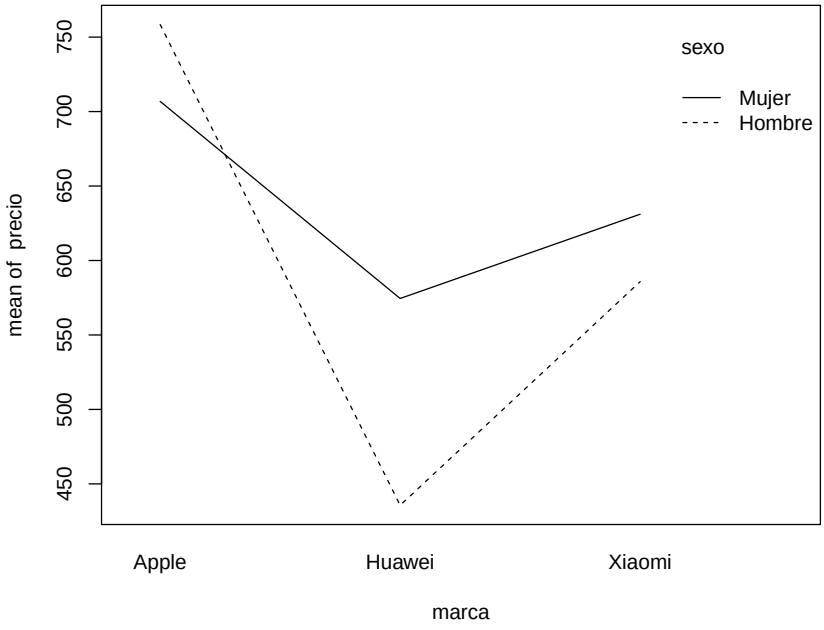
También podemos obtener estadísticos descriptivos de la respuesta para cada combinación de niveles de los factores:

```
aggregate(precio ~ sexo + marca, movil, mean)
```

```
##     sexo  marca   precio
## 1 Hombre  Apple 758.4304
## 2  Mujer  Apple 706.8320
## 3 Hombre Huawei 435.6322
## 4  Mujer Huawei 574.4963
## 5 Hombre Xiaomi 585.8118
## 6  Mujer Xiaomi 631.0839
```

También es habitual representar gráficamente estos estadísticos:

```
with(movil, interaction.plot(marca, sexo, precio))
```



En el caso de múltiples variables categóricas podemos generar tablas de contingencia multidimensionales y, por ejemplo, gráficos de barras agrupados con las frecuencias de dos variables categóricas para cada combinación de categorías de las demás. Por ejemplo:

```
# Distribución conjunta de sexo, nsatisfa y marca
frec <- with(movil, table(sexo, nsatisfa, marca))
frec
```

```
## , , marca = Apple
##
##         nsatisfa
## sexo     Bajo Medio Alto
##   Hombre    2     2    2
##   Mujer     0     3    3
##
## , , marca = Huawei
##
##         nsatisfa
## sexo     Bajo Medio Alto
##   Hombre    4     1    5
##   Mujer     3     3    2
##
## , , marca = Xiaomi
##
##         nsatisfa
## sexo     Bajo Medio Alto
##   Hombre    1     3    4
##   Mujer     4     4    4
```

```
# Distribución de nsatisfa condicionada a sexo y marca:
porc.cond <- 100*prop.table(frec, c(1,3))
```

```
round(porc.cond, 1)
```

```
## , , marca = Apple
##
##         nsatisfa
## sexo     Bajo Medio Alto
##   Hombre 33.3  33.3 33.3
##   Mujer   0.0  50.0 50.0
##
## , , marca = Huawei
##
##         nsatisfa
## sexo     Bajo Medio Alto
##   Hombre 40.0  10.0 50.0
##   Mujer  37.5  37.5 25.0
##
## , , marca = Xiaomi
##
##         nsatisfa
## sexo     Bajo Medio Alto
##   Hombre 12.5  37.5 50.0
##   Mujer  33.3  33.3 33.3
```