Análisis Estadístico Multivariante

Francisco Javier Mercader Martínez

Índice

Tema 1: Vectores aleatorios

1.1) Introducción

Objetivo: estudiar k variables sobre una población de individuos (objetos).

Algunos ejemplos:

→ Las variables meteorológicas como temperatura, humedad y velocidad del viento.

→ La intensidad y la fase de una señal aleatoria que se miden en los canales de comunicación.

→ Los parámetros clínicos de los pacientes (como presión arterial, niveles de glucosa, etc.)

Habitualmente estas variables cualitativas o discretas que nos indicarán grupos de individuos.

Estas variables se representarán mediante vectores aleatorios sobre un espacio de probabilidad.

1) Definiciones

Un vector aleatorio (v.a.) k-dimensional sobre un espacio de probabilidad $(\Omega, \mathcal{S}, \mathcal{P})$ es $X = (X_1, \ldots, X_k)$ tal que

$$X_i^{-1}(-\infty, x] \in \mathcal{S}$$

para todo $x \in \mathbb{R}$, $i = 1, \ldots, k$

- Función de distribución conjunta

$F : \mathbb{R}^k \longrightarrow [0, 1]$,

$$F(x_1, \ldots, x_k) \coloneqq P[X_1 \leq x_1, X_2 \leq x_2, \ldots, X_k \leq x_k],$$

para todo $x_1, \ldots, x_k \in \mathbb{R}$.

1.2) Independencia de las variables aleatorias

- Definición

Las variables aleatorias $X_1, \ldots, X_k$ son independientes si los sucesos

$$\{x_1 \leq x_1\}, \{X_2 \leq x_2\}, \ldots, \{X_k \leq x_k\}$$

son independientes para todo $x_1, \ldots, x_k \in \mathbb{R}$.

Esto es equivalente a que

$$F(x_1, \ldots, x_k) = P[X_1 \leq x_1] \cdot P[X_2 \leq x_2] \cdots P[X_k \leq x_k]$$

para todo $x_1, \ldots, x_k \in \mathbb{R}$.

1.3) Distribuciones marginales

La función $F_{X_i}(x_i) = P[X_i \leq x_i]$ se denomina función de distribución marginal i-ésima y corresponde con la función de distribución de la variable aleatoria X_i

Las distribuciones marginales pueden obtenerse a partir de la distribución conjunta:

$$F_{X_i}(x_I) = F(+\infty, \ldots, +\infty, x_i, +\infty, \ldots, +\infty)$$

Análogamente, la función de distribución marginal del subvector aleatorio $(X_{i_1}, \ldots, X_{i_m})$ vendrá dada por

$$F_{X_{i_1},\ldots,X_{i_m}}(x_{i_1}, \ldots, x_{i_m}) = F(+\infty, \ldots, +\infty, x_{i_1}, +\infty, \ldots, +\infty, x_{i_m}, +\infty, \ldots, +\infty).$$

1.4) Vector aleatorio absolutamente continuo

Un vector aleatorio X es absolutamente continuo si existe una función $f : \mathbb{R}^k \longrightarrow \mathbb{R}$ no negativa (llamada función de densidad) tal que

$$F(x) = F(x_1, \ldots, x_k) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_k} f(z_1, \ldots, z_k) \mathrm{d}z_k, \ldots, \mathrm{d}z_1,$$

para todo $x = (x_1, \ldots, x_k) \in \mathbb{R}^k$

Usando el teorema fundamental del cálculo, se tiene que en cada punto de continuidad $(x_1, \ldots, x_k)$ de f:

$$\frac{\partial^k F(x_1, \ldots, x_k)}{\partial x_1, \ldots, \partial x_k} = f(x_1, \ldots, x_k).$$

Existen variables aleatorias cuya función de distribución es continua pero que no son absolutamente continuas (tienen una parte singular) y puede ocurrir que $X_1, \ldots, X_k$ sean absolutamente continuas y que $(X_1, \ldots, X_k)$ no lo sea.

→ Ejemplo: Si X_1 es una variable aleatoria absolutamente continua, entonces el vector aleatorio $X = (X_1, X_2)$ es continuo pero no absolutamente continuo.

→ De hecho, es completamente singular ya que está contenido en la recta $y = x$ que tiene medida cero en $\mathbb{R}^2$.

Esto ocurre si consideramos las notas de unos alumnos y sus medidas. En estos casos deberemos eliminar estas variables dependientes del vector.

1.5) Vector aleatorio discreto

Un vector aleatorio X se dice que es discreto si existe un conjunto numerable $\mathcal{S} \in \mathbb{R}^k$ tal que $P(X \in \mathcal{S}) = 1$.

Función masa de probabilidad de una vector aleatorio discreto:

$$P[X = x] = P[X_1 = x_1, \ldots, X_k = x_k]$$

para todo $x = (x_1, \ldots, x_k) \in \mathbb{R}^k$, satisfaciendo:

→ $P[X = x] \geq 0, \ \forall x \in \mathcal{S}$

→ $\sum_{x \in \mathcal{S}} P[X = x] = 1$

Función de distribución de un vector aleatorio discreto:

$$F(x) = P[X \leq x] = \sum_{\substack{z \in \mathcal{S} \\ z \leq x}} P[X = z],$$

para todo $x \in \mathbb{R}^k$.

1.6) Distribuciones marginales

1.6.1) Caso continuo

- Distribución marginal de la variable aleatoria X_i

Sea $X = (X_1, \ldots, X_k)$ un vector aleatorio continuo con función de densidad f entonces cada componente X_i es de tipo continuo y su función de distribución es;

$$F_{X_i}(x_i) = P[X_i \leq x_i] = \int_{-\infty}^{x_i} f_{X_i}(z_i)\mathrm{d}z_i,$$

con

$$f_{X_i} = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f(z_1, \ldots, z_k)\mathrm{d}z_1, \ldots, \mathrm{d}z_{i-1} \cdot \mathrm{d}z_{i+1} \ldots, \mathrm{d}z_k,$$

para todo $z_i \in \mathbb{R}$.

La función de densidad marginal de cualquier subvector se calcularía de igual forma.

$X_1, \ldots, X_k$ son independientes $\longleftrightarrow f(x_1, \ldots, x_k) = f_{X_1}(x_1) \cdots f_{X_k}(x_k)$.

1.6.2) Caso discreto

- Distribución marginal de la variable aleatoria X_i

Sea $X = (X_1, \ldots, X_l)$ un vector aleatorio discreto con $P[X \in \mathcal{S}] = 1$ y función masa de probabilidad $P[X = x]$, para todo $x \in \mathcal{S}$.

Si X_i es una componente arbitraria y por tanto discreta con valores en $\mathcal{S}_i$, entonces su función masa de probabilidad puede obtenerse a partir de la conjunta:

$$P[X_i = x_i] = \sum_{\substack{x_1 \ldots, x_{i-1}, x_{i+1}, \ldots, x_k \\ (x_1, \ldots, x_i, \ldots, x_n) \in \mathcal{S}}} P[X_1 = x_1, \ldots, X_{i-1} = x_{i-1}, X_i = x_i, X_{i+1} = x_{i+1}, \ldots, X_k = x_k].$$

La función masa de probabilidad marginal de cualquier subvector se calcularía de igual forma.

$X_1, \ldots, X_k$ son independientes $\longleftrightarrow$ para todo $(x_1, \ldots, x_k) \in \mathcal{S}$,

$$P[X_1 = x_1, \ldots, X_k = x_k] = P[X_1 = x_1] \cdots P[X_k = x_k].$$

Nota:
A y B independientes $\longleftrightarrow P(A|B) = P(A)$
$P(A|B) = \dfrac{P(A \cap B)}{P(B)} = P(A) \longrightarrow P(A \cap B) = P(A) \cdot P(B)$

1.7) Distribuciones condicionadas

1.7.1) Caso continuo

- Distribución condicionada al valor de una variable

Sea $X = (X_1, \ldots, X_k)$ un vector aleatorio continuo con función de densidad f.
Sea X_i una componente arbitraria y $x_i^* \in \mathbb{R}$ tal que $f_{X_i}(x_i^*) > 0$.
Se define la distribución condicionada de $(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_k)$ a $(X_i = x_i^*)$ como la determinada por la función de densidad:

$$f_{X_1, \ldots, X_{i-1}, \ldots, X_k | X_i = x_i^*}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k | x_i^*) = \frac{f(x_1, \ldots, x^*, \ldots, x_k)}{f_{X_i}(x_i^*)}.$$

- Distribución condicionada a valores de varias variables

Sea $X = (X_1, \ldots, X_k)$ un vector aleatorio continuo con función de densidad f.
Sea $(X_{i_1}, \ldots, X_{i_m})$ un subvector arbitrario y $(x^*_{i_1}, \ldots, x^*_{i_m}) \in \mathbb{R}^m$ tal que:

$$f_{X_{i_1},\ldots,X_{i_m}}(x^*_{i_1}, \ldots, x^*_{i_m}) > 0.$$

Se define la distribución condicionada de $(X_1, \ldots, X_{i_1-1}, X_{i_1+1}, \ldots, X_{i_m-1}; X_{i_m+1}, \ldots, X_k)$ a $(X_{i_1} = x^*_{i_1}, \ldots, X_{i_m} = x_{i^*_m})$ como la determinada por la función de densidad:

$$f_{X_1,\ldots,X_{i_1-1},X_{i_1+1},\ldots,X_{i_m-1},\ldots,X_k|X_{i_1}=x^*_{i_1},\ldots,X_{i_m}=x_{i^*_m}}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{i_m-1}, x_{i_m+1}, \ldots, x_k|x^*_i) = \frac{f(x_1, \ldots, x^*_{i_1}, \ldots, x^*_{i_m} \ldots, x_k)}{f_{X_{i_1},\ldots,X_{i_m}}(x^*_{i_1}, \ldots, x^*_{i_m})}$$

1.7.2) Caso discreto

- Distribución condicionada al valor de una variable

Sea $X = (X_1, \ldots, X_k)$ un vector aleatorio discreto.
Sea X_i una componente arbitraria y $x^*_i \in \mathbb{R}$ tal que

$$P[X_i = x^*_i] > 0.$$

Se define la distribución condicionada de $(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_k)$ a $(X_i = x^*_i)$ como la determinada por la función masa de probabilidad:

$$P[X_1 = x_1, \ldots, X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}, \ldots, X_k = x_k|X_i = x^*_i] = \frac{P[X_1 = x_1, \ldots, X_{i-1} = x_{i-1}, X_i = x^*_i, X_{i+1} = x_{i+1}, \ldots, X_k = x_k]}{P[X_i = x^*_i]}$$

para todo $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$ tal que $x_1, \ldots, x_{i-1}, x^*_i, x_{i+1}, \ldots, x_k \in \mathcal{S}$.

- Distribución condicionada a valores de varias variables

Sea $X = (X_1, \ldots, X_k)$ un vector aleatorio discreto.

Sea $X_{i_1}, \ldots, X_{i_m}$ un subvector arbitrario y $(x^*_{i_1}, \ldots, x^*_{i_m}) \in \mathbb{R}^m$ tal que

$$P[X_{i_1} = x^*_{i_1}, \ldots, X_{i_m} = x^*_{i_m}] > 0.$$

Se define la distribución condicionada de $(X_1, \ldots, X_{i_1-1}, X_{i_1+1}, \ldots, X_{i_m-1}, X_{i_m+1}, \ldots, X_k)$ a $(X_{i_1} = x^*_{i_1}, \ldots, X_{i_m} = x^*_{i_m})$ como la determinada por la función masa de probabilidad:

$$P[X_1 = x_1, \ldots, X_{i_1-1} = x_{i_1-1}, X_{i_1+1} = x_{i_1+1}, \ldots, X_{i_m-1} = x_{i_m-1}, X_{i_m+1} = x_{i_m+1}, \ldots, X_k = x_k|X_{i_1} = x^*_{i_1}, \ldots, X_{i_m} = x_{i^*_m}] = \frac{P[X_1 = x_1, \ldots, X_{i_1} = x^*_{i_1}, \ldots, X_{i_m} = x^*_{i_m}, \ldots, X_k = x_k]}{P[X_{i_1} = x^*_{i_1}, \ldots, X_{i_m} = x^*_{i_m}]}$$

para todo $(x_1, \ldots, x_{i_1}, x_{i_1+1}, \ldots, x_{i_m-1}, x_{i_m+1}, \ldots, x_k)$, tal que $(x_1, \ldots, x^*_{i_1}, \ldots, x^*_{i_m}, \ldots, x_k) \in \mathcal{S}$

1.8) Distribución normal multivariante $\mathcal{N}_k(\mu, V)$

1) Función de densidad

$$f(x) = \frac{1}{\sqrt{|V|(2\pi)^k}} \exp\left(-\frac{1}{2}(x-\mu)'V^{-1}(x-\mu)\right),$$

para $x \in \mathbb{R}^k$, donde μ es un vector k-dimensional y V es una matriz $k \times k$ simétrica y definida positiva.

- Definiciones

Una matriz simétrica A, de dimensión $k \times k$, se dice que es definida positiva si se verifica que $x'Ax > 0$ para cualquier vector no nulo $x \in \mathbb{R}^k$.

Una matriz simétrica A, de dimensión $k \times k$, se dice que es semidefinida positiva si se verifica que $x'Ax \geq 0$ para cualquier vector $x \in \mathbb{R}^k$.

¿Cómo calcular la inversa de $V = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}$ con R?

```
V <- matrix(c(1, 1/2,
              1/2, 1), nrow = 2, ncol = 2, byrow = TRUE)
solve(V)
```

```
##           [,1]       [,2]
## [1,]  1.3333333 -0.6666667
## [2,] -0.6666667  1.3333333
```

1.8.1) Normal bivariante

- Función de densidad

Caso bivariante, $k = 2$, para $\mu = (0, 0)$ y $V = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}$.

Cálculo de la función de densidad en $x = (1, 1)$ utilizando la función `dmvnorm` de la librería `mvtnorm` de R:

```
library("mvtnorm")
V <- matrix(c(1, 1/2,
                1/2, 1), nrow = 2, ncol = 2, byrow = TRUE)
mu <- c(0, 0)
x <- c(1, 1)
dmvnorm(x, mean = mu, sigma = V)
```

```
## [1] 0.0943539
```

- Función de distribución

Cálculo (aproximado) de la función de distribución en $x = (1, 1)$ con la función:

`pmvnorm(lower = -Inf, upper = x, mean = mu, sigma = V)`

```
library("mvtnorm")
V <- matrix(c(1, 1/2,
                1/2, 1), nrow = 2, ncol = 2, byrow = TRUE)
mu <- c(0, 0)
x <- c(1, 1)
pmvnorm(lower = -Inf, upper = x, mean = mu, sigma = V)
```

```
## [1] 0.7452036
```

- Probabilidad en rectángulos

Cálculo (aproximado) de las probabilidades en rectángulos dando los límites inferiores y superiores del rectángulo. Por ejemplo, para calcular

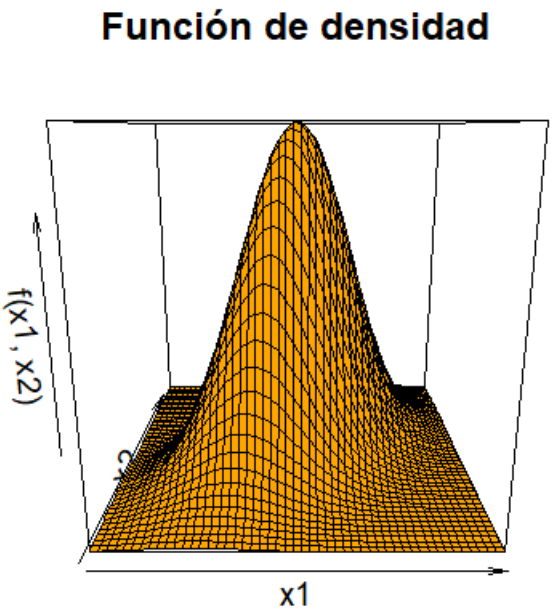
$$P(-1 < X_1 < 1,\ -1 < X_2 < 1)$$

```r
library("mvtnorm")
V <- matrix(c(1, 1/2,
              1/2, 1), nrow = 2, ncol = 2, byrow = TRUE)
mu <- c(0, 0)
x1 <- c(- 1, -1)
x2 <- c(1, 1)
pmvnorm(lower = x1, upper = x2, mean = mu, sigma = V)
```

```
## [1] 0.499718
```

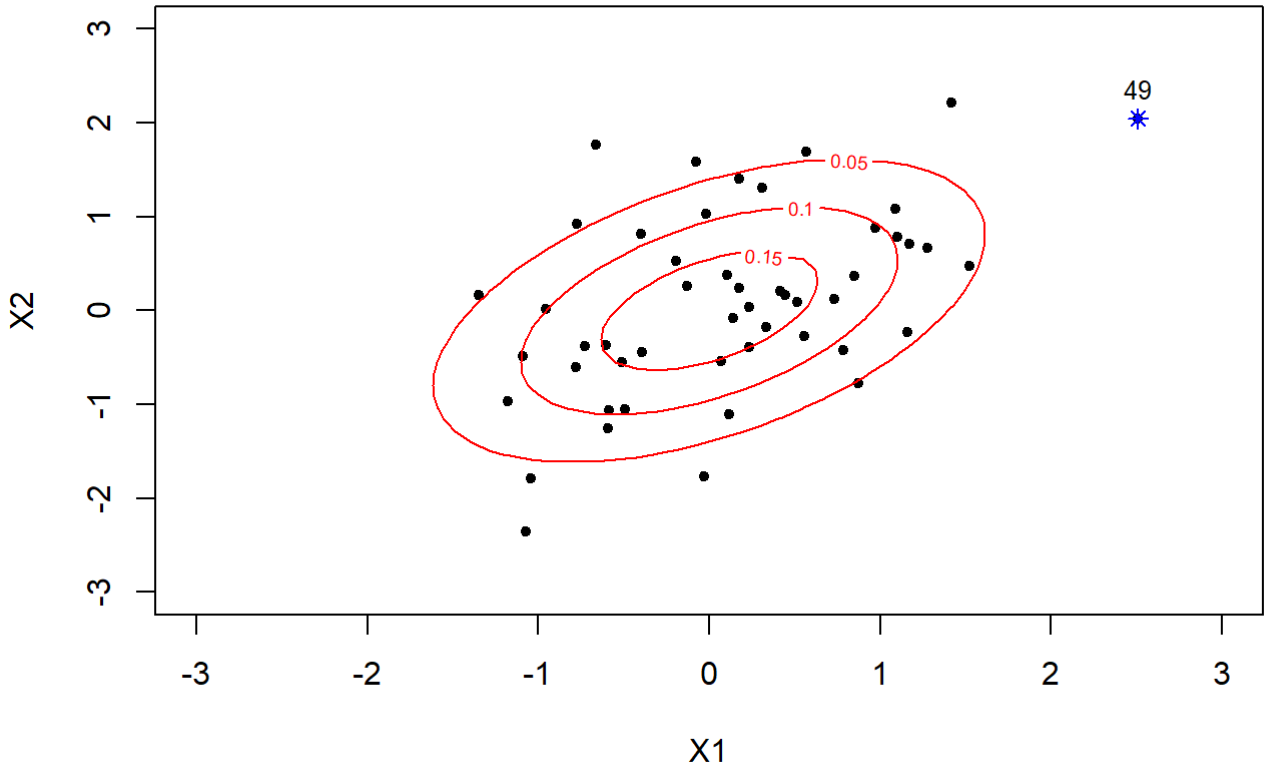
Su representación gráfica:

```r
f <- function(x1, x2) dmvnorm(data.frame(x1, x2), mu, V)
x <- seq(-3, 3, length = 50)
y <- seq(-3, 3, length = 50)
z <- outer(x, y, f)
persp(x, y, z, xlab = 'x1', ylab = 'x2', zlab = 'f(x1, x2)', col = 'orange', main = "
    Función de densidad")
```



Su representación gráfica ($f(x_1, x_2) = c$) y 50 datos simulados de este modelo

```r
#Se fija la semilla para la generación aleatoria
set.seed(123)
#Generación aleatoria del modelo
d <- rmvnorm(50, mu, V)
plot(d, xlab = "X1", ylab = "X2", pch = 20, xlim = c(-3, 3), ylim = c(-3, 3))
contour(x, y, z, nlevels = 4, add = T, col = 'red')
```



- Distancia de Mahalanobis

La distancia de Mahalanobis del vector x al vector μ basada en la matriz V:

$$D = \sqrt{(x-\mu)'V^{-1}(x-\mu)}$$

Tiene en cuenta la diferentes escalas de los datos y sus correlaciones.

Servirá para detectar las observaciones más alejadas del vector de medias que podrían ser observaciones atípicas (`outliers`) que no provengan de nuestra población o contengan errores.

→ Cuando se pueda, se deberán chequear y, si es posible, corregir o eliminar.

→ En otros casos, se deberán mantener por ser observaciones correctas que hay que tener en cuenta.

- Cálculo de la distancia de Mahalanobis

Para calcular las distancias de Mahalanobis al cuadrado de los datos al vector de medias (teóricas o muestrales) podemos utilizar la función `mahalanobis`.

```
V <- matrix(c(1, 1/2,
              1/2, 1), nrow = 2, ncol = 2, byrow = TRUE)
mu <- c(0, 0)
dM1 <- mahalanobis(d, mu, V)
dM2 <- mahalanobis(d, colMeans(d), cov(d))
```

- Distancias de los datos simulados al vector de medias teóricas μ con respecto a V

```
summary(dM1)
```

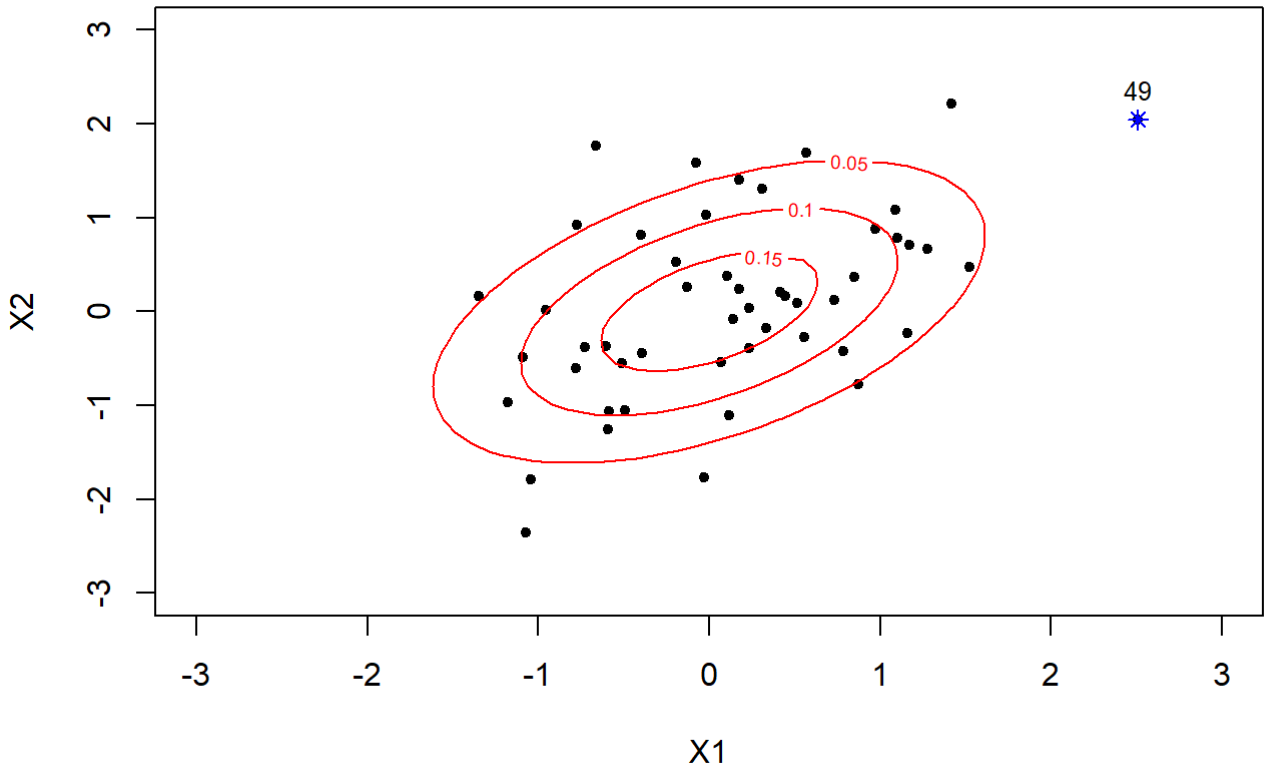
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.05216 0.41016 1.26433 1.66615 2.31591 7.13332
```

- ¿Dónde se encuentra la observación más alejada del vector de medias?

```
d[which.max(dM1), ]
```

```
## [1] 2.509470 2.046512
```

```
f <- function(x1, x2) dmvnorm(data.frame(x1, x2), mu, V)
x <- seq(-3, 3, length = 50)
y <- seq(-3, 3, length = 50)
z <- outer(x, y, f)
plot(d, xlab = "X1", ylab = "X2", pch = 20, xlim = c(-3, 3), ylim = c(-3, 3))
points(d[which.max(dM1), 1], d[which.max(dM1), 2], col = "blue", pch = 8)
text(d[which.max(dM1), 1], d[which.max(dM1), 2], which.max(dM1), cex = 0.8, pos = 3)
contour(x, y, z, nlevels = 4, add = T, col = 'red')
```



- Distancias de los datos simulados al vector de medias muestrales $\overline{x}$ con respecto a $\mathcal{S}$

```
summary(dM2)
```

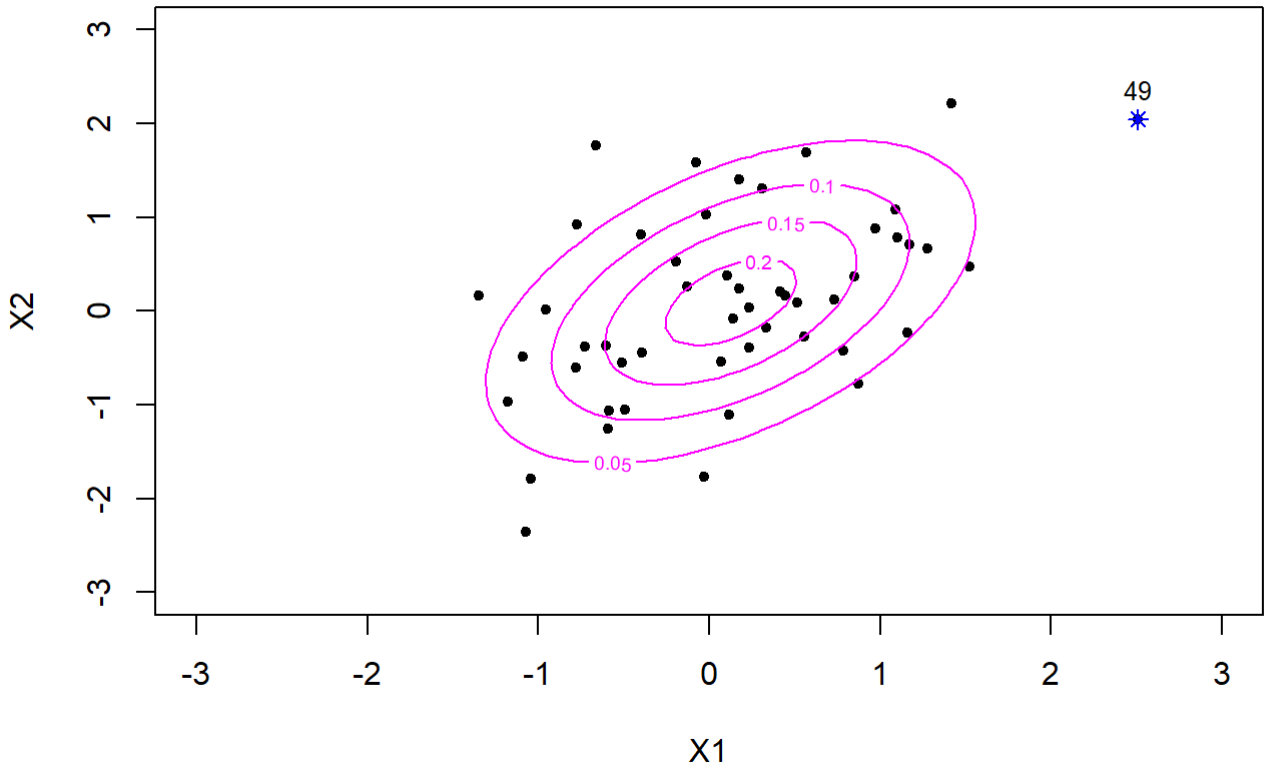
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.02114 0.67111 1.52636 1.96000 2.64131 8.65906
```

- ¿Dónde se encuentra la observación más alejada del vector de medias?

```
d[which.max(dM2), ]
```

```
## [1] 2.509470 2.046512
```

```r
f <- function(x1, x2) dmvnorm(data.frame(x1, x2), colMeans(d), cov(d))
x <- seq(-3, 3, length = 50)
y <- seq(-3, 3, length = 50)
z <- outer(x, y, f)
plot(d, xlab = "X1", ylab = "X2", pch = 20, xlim = c(-3, 3), ylim = c(-3, 3))
points(d[which.max(dM2), 1], d[which.max(dM2), 2], col = "blue", pch = 8)
text(d[which.max(dM2), 1], d[which.max(dM2), 2], which.max(dM2), cex = 0.8, pos = 3)
contour(x, y, z, nlevels = 4, add = T, col = 'magenta')
```



1.9) Distribución multinomial $\mathcal{M}_k(n, p_1, \ldots, p_k)$

- Modelo multinomial

$(X_1, \ldots, X_k)$: variables aleatorias que representan el número de veces que ocurre el suceso A_i en un experimento aleatorio repetido n veces con k opciones dadas por $\{A_1, \ldots, A_k\}$ y con probabilidades constantes $p_i = P(A_i)$, para $i = 1, \ldots, k$.
Función masa de probabilidad conjunta:

$$p(x_1, \ldots, x_k) = P[X_1 = x_1, \ldots, X_k = x_k] = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k},$$

para enteros no negativos tales que $x_1 + \cdots + x_k = n$ y donde $p_i \in [0, 1]$ satisface $p_1 + \cdots + p_k = 1$.

Distribuciones marginales: X_i sigue una distribución binomial $B(n, p_i)$, con $E(X_i) = np_i$.

1.10) Estadístico de Pearson

- Discrepancias entre lo observado y lo esperado

Contexto: Lanzamos un dado $n =$ veces, $p_i = \dfrac{1}{6}$ para todo i, y los valores esperados son $np_i = 10$, para $i = 1, \ldots, 6$.

Objetivo: Medir las discrepancias entre valores observados y esperados.

Sea $X = (X_1, \ldots, X_k)$ una variable aleatoria con distribución multinomial, entonces el estadístico

$$T = \sum_{i=1}^{k} \frac{X_i - np_i}{np_i}$$

sigue una distribución Chi-cuadrado χ^2_{k-1} de Pearson con $k-1$ grados de libertad, cuando $n \longrightarrow \infty$.

1.11) Medias y covarianzas

- Definiciones

Dado el vector aleatorio.

→ El vector de medias (o esperanza matemática de X) se define como:

$$\mu \coloneqq E[X] = (E[X_1], \ldots, E[X_k])' = (\mu_1, \ldots, \mu_k)'$$

(note que es un vector columna).

→ La matriz de covarianzas (o varianzas-covarianzas) se define como:

$$V = (\sigma_{i,j}),$$

donde $\sigma_{i,j}$ es la covarianza entre X_i y X_j, definida como:

$$\sigma_{i,j} = \mathrm{Cov}(X_i, X_j) = E\left[(X_i - \mu_i)(X_j - \mu_j)\right] = E[X_i X_j] - \mu_i \mu_j$$

Notemos que $\sigma_{i,i} = E\left[(X_i - \mu_i)^2\right] = \mathrm{Var}(X_i) = \sigma_i^2$.

- Cálculo de la esperanza matemática

La media de cada componente X_i del vector puede calcularse a partir de la distribución conjunta o a partir de la marginal.

→ Caso discreto:

$$\begin{aligned} E[X_i] &= \sum_{x_i} x_i P[X_i = x_i] \\ &= \sum_{x_1, \ldots, x_k} x_i P[X_1 = x_1, \ldots, X_k = x_k] \end{aligned}$$

→ Caso continuo:

$$\begin{aligned} E[X_i] &= \int_{\mathbb{R}} x_i f_{X_i}(x_i)\, \mathrm{d}x_i \\ &= \int_{\mathbb{R}^k} x_i f(x_1, \ldots, x_k)\, \mathrm{d}x_1 \cdots \mathrm{d}x_k \end{aligned}$$

1.11.1) Esperanza de la transformación $g : \mathbb{R}^k \longrightarrow \mathbb{R}$

- Caso discreto

Sea $g : \mathbb{R}^k \longrightarrow \mathbb{R}$ una función medible $\longrightarrow Y = g(X)$ es una variable aleatoria .

Si X es de tipo discreto,

$$\exists E[g(X)] \longleftrightarrow \sum_{x_1,\ldots,x_k} |g(x_1,\ldots,x_k)|\, P[X_1 = x_1,\ldots,X_k = x_k] < \infty$$

Y en caso de existir:

$$E[g(X_1,\ldots,X_k)] = \sum_{x_1,\ldots,x_k} g(x_1,\ldots,x_k) P[X_1 = x_1,\ldots,X_k = x_k]$$

- Caso continuo

Sea $g : \mathbb{R}^k \longrightarrow \mathbb{R}$ una función medible $\longrightarrow Y = g(X)$ es una variable aleatoria .

Si X es de tipo continuo,

$$\exists E[g(X)] \longleftrightarrow \int_{\mathbb{R}^k} |g(x_1,\ldots,x_k)|\, f_X(x_1,\ldots,x_k)\, \mathrm{d}x_1 \cdots \mathrm{d}x_k < \infty$$

Y en caso de existir:

$$E[g(X_1,\ldots,X_k)] = \int_{\mathbb{R}^k} g(x_1,\ldots,x_k) f_X(x_1,\ldots,x_k)\, \mathrm{d}x_1 \cdots \mathrm{d}x_k.$$

- Propiedades

V es una matriz simétrica y semidefinida positiva ($x'Vx \geq 0$, para todo $x \in \mathbb{R}^k$).

En forma matricial,

$$V = E\left[(X-\mu)(X-\mu)'\right] = E[XX'] - \mu\mu'.$$

donde la esperanza de una matriz aleatoria se define como la matriz de las esperanzas de cada variable.

Si X_i y X_j son independientes, entonces

$$E[X_i X_j] = E[X_i]E[X_j]$$

y, por lo tanto, $\mathrm{Cov}(X_i, X_j) = 0$. El recíproco no es cierto.

Si $X \longrightarrow \mathcal{N}_k(\mu, V)$, se puede demostrar que μ es el vector de medias y V es la matriz de covarianzas.

1.12) Correlación

La correlación (lineal de Pearson) entre X_i y X_j se define como

$$\rho_{i,j} = \mathrm{Corr}(X_i, X_j) = \frac{\sigma_{i,j}}{\sigma_i \sigma_j}$$

siendo $\rho_{i,i} = \mathrm{Corr}(X_i, X_j) = 1$.

Mide el grado de relación lineal entre X_i y X_j.

Puede demostrarse que

$$-1 \leq \rho_{i,j} \leq 1.$$

Se dice que X_i y X_j son incorreladas si $\rho_{i,j} = 0$.

Si son independientes serán incorreladas, pero el recíproco no es cierto.

La matriz de correlaciones es $R = (\rho_{i,j})$.

1.12.1) Correlación entre vectores aleatorios

Análogamente, si X e Y son vectores aleatorios (de dimensiones cualesquiera), se define su matriz de covarianzas como

$$\mathrm{Cov}(X,Y) = (\mathrm{Cov}(X_i, Y_j))$$

y su matriz de correlaciones como

$$\mathrm{Corr}(X,Y) = (\mathrm{Corr}(X_i, Y_j)).$$

Puede demostrarse que

$$\mathrm{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])'].$$

Evidentemente, $\mathrm{Cov}(X) = \mathrm{Cov}(X,X)$.

- Propiedades

Si X, Y, Z son vectores (columna) aleatorios, se verifican las propiedades siguientes:

1) $E[a_1 g_1(X) + a_2 g_2(X)] = a_1 E[g_1(X)] + a_2 E[X_2]$, donde $a_1, a_2 \in \mathbb{R}$ y g_1 y g_2 son funciones medible de vectores aleatorios.

2) $X = (Y,Z), E_X[g(Y)] = E_Y[g(Y)]$, donde g es una función medible de un vector aleatorio, E_X denota la esperanza en la distribución conjunta y E_Y en la distribución marginal.

3) Si X e Y son independientes, entonces

$$E[g_1(X)g_2(Y)] = E[g_1(X)]E[g_2(Y)],$$

donde g_1 y g_2 son funciones medibles cualesquiera de vectores aleatorios .

4) $E[AX + b] = AE[X] + b, A \in M_{m,k} b' \in \mathbb{R}^m$.

5) $\mathrm{Cov}(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j]$.

6) Si $X_1, \ldots, X_k$ son independientes, $\mathrm{Cov}(X_i, X_j) = 0$.

7) $\mathrm{Var}(X_i + X_j) = \mathrm{Var}(X_i) + 2\,\mathrm{Cov}(X_i, X_j) + \mathrm{Var}(X_j)$.

8) $\mathrm{Cov}(aX_i + b, cX_j + d) = ac\,\mathrm{Cov}(X_i, X_j)$, donde $a, b, c, d \in \mathbb{R}$.

9) $\mathrm{Cov}(X) = E[(X - \mu)(X - \mu)'] = E[XX'] - \mu\mu'$.

10) $\mathrm{Var}(a'X) = a'\,\mathrm{Cov}(X)a = \sum_{i,j} a_i a_j \sigma_{i,j}$, donde $a \in \mathbb{R}^k$.

11) $\mathrm{Cov}(AX + b) = A\,\mathrm{Cov}(X)A'$, donde $A \in M_{m,k}$ y $b' \in \mathbb{R}^m$.

12) Si $X_1, \ldots, X_k$ son independientes, $\mathrm{Corr}(X_i, X_j) = 0$.

13) $\mathrm{Corr}(aX_i + b, cX_j + d) = \mathrm{Corr}(X_i, X_j)$, donde $a, b, c, d \in \mathbb{R}$.

14) $-1 \leq \mathrm{Corr}(X_i, X_j) \leq 1$.

15) $\mathrm{Corr}(X_i, aX_i + b) = \pm 1$, donde $a, b \in \mathbb{R}$ (según el signo de a).

16) $\mathrm{Corr}(X) = \delta^{-1}\,\mathrm{Cov}(X)\delta^{-1}$, donde δ es la matriz diagonal formada por las desviaciones típicas ($\delta = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$).

17) $\mathrm{Cov}(X,Y) = (\mathrm{Cov}(X_i, Y_j)) = \mathrm{Cov}(Y,X)'$.

18) $\mathrm{Cov}(X + Z, Y) = \mathrm{Cov}(X,Y) + \mathrm{Cov}(Z,Y)$

19) Si X e Y tienen la misma dimensión, entonces $\mathrm{Cov}(X + Y) = \mathrm{Cov}(X) + \mathrm{Cov}(X,Y) + \mathrm{Cov}(Y,X) + \mathrm{Cov}(Y)$.

20) $\mathrm{Cov}(AX, BY) = A\,\mathrm{Cov}(X,Y)B'$, donde A y B son matrices (de dimensiones adecuadas).

21) Si X e Y independientes, entonces $\mathrm{Cov}(X,Y) = 0$.

- Demostración apartado (10)

Directamente se tiene que:

$$\mathrm{Var}(a'X) = \mathrm{Cov}(a'X, a'X) = E[a'(X-\mu)(X-\mu)'a] = a\,\mathrm{Cov}(X)a$$

Como consecuencia, se obtiene que la matriz de covarianzas $\mathrm{Cov}(X)$ es semidefinida positiva ya que $\mathrm{Var}(a'X) \geq 0$.

Lo mismo le ocurre a la matriz de correlaciones $\mathrm{Corr}(X)$ ya que es la matriz de covarianzas de las variables aleatorias tipificadas $Z_i = \dfrac{X_i - \mu_i}{\sigma_i}$.

1.13) Resultados básicos de la inferencia

- Contexto

En la práctica, todas las medidas, varianzas y covarianzas serán desconocidas por lo que tenemos que estimarlas.
Para ello dispondremos de una muestra de individuos (objetos) en los que se han medido todas las variables.
Proporcionamos resultados básicos de inferencia para poder aplicar las técnicas multivariantes que desarrollaremos en temas posteriores.
Se ilustran estos procedimientos de inferencia con conjuntos de datos de `R`, accesibles con `data()`.

1.13.1) ¿Cómo se representan las muestras aleatorias?

- Matriz de la muestra aleatoria simple

En general, nuestra muestra aleatoria se representará como:

i	X_1	X_2	$\cdots$	X_k	Y
$\mathbf{O}_1$	$X_{1,1}$	$X_{1,2}$	$\cdots$	$X_{1,k}$	Y_1
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\mathbf{O}_i$	$X_{i,1}$	$X_{i,2}$	$\cdots$	$X_{i,k}$	Y_i
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\mathbf{O}_n$	$X_{n,1}$	$X_{n,2}$	$\cdots$	$X_{n,k}$	Y_n

La variable Y solo se usará para detonar la variable respuesta en regresión.

Si no hay grupos supondremos que los objetos

$$O_i = (X_{i,1}, X_{i,2}, \ldots, X_{i,k})'$$

son una muestra aleatoria simple de X, es decir, serán vectores (columna) aleatorios independientes con la misma distribución que X.
Si no hay grupos supondremos lo mismo en cada grupo

En algunos casos usaremos la matriz $M = (X_{i,j})$ que será una matriz aleatoria.

1.13.2) ¿Cómo se muestran los valores muestrales?

- Matriz de datos

En general, nuestra muestra se representará como:

i	x_1	x_2	$\cdots$	x_k	y
$\mathbf{o}_1$	$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,k}$	y_1
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\mathbf{o}_i$	$x_{i,1}$	$x_{i,2}$	$\cdots$	$x_{i,k}$	y_i
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\mathbf{o}_n$	$x_{n,1}$	$x_{n,2}$	$\cdots$	$x_{n,k}$	y_n

Si no hay grupos, supondremos que los vectores

$$o_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,k})'$$

son una realización de una muestra aleatoria simple de X, es decir, serán vectores (columna) con los datos muestrales.
Si hay grupos supondremos lo mismo en cada grupo.

La variable Y solo se usará para detonar la variable respuesta en regresión.

En algunos casos usaremos la matriz de datos $M = (x_{i,j})$

1.13.3) El conjunto de datos `LifeCycleSavings`

- Cargamos los datos y visualizamos las primeras filas

```
datos <- LifeCycleSavings
head(datos, n = 6)
```

```
##              sr pop15 pop75     dpi ddpi
## Australia 11.43 29.35  2.87 2329.68 2.87
## Austria   12.07 23.32  4.41 1507.99 3.93
## Belgium   13.17 23.80  4.43 2108.47 3.82
## Bolivia    5.75 41.89  1.67  189.13 0.22
## Brazil    12.88 42.19  0.83  728.47 4.56
## Canada     8.79 31.72  2.85 2982.88 2.43
```

- ¿Qué información está recogida en el conjunto de datos?

Con la instrucción `help(LifeCycleSavings)` conocemos qué información está contenida en el conjunto:

- `sr`: incremento de los ahorros personales 1960-1970.
- `pop15`: % población menor de 15 años.
- `pop75`: % población menor de 75.
- `dpi`: ingresos per-capita.

1.14) Estimador para el vector de medias μ

Vector de medias muestrales, también llamado objeto medio, se define como:

$$\overline{O} = \overline{X} = (\overline{X}_1, \ldots, \overline{X}_k)' = \frac{1}{n}\sum_{i=1}^{n} O_i,$$

donde $\overline{X}_j = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_{i,j}$.

Se puede demostrar fácilmente que:

$\rightarrow$ $E(\overline{O}) = \mu$ (estimador centrado de μ)

$\rightarrow$ $\mathrm{Cov}(\overline{O}) = \dfrac{V}{n}$

1.14.1) ¿Dónde se encuentra el vector de medias muestrales?

- Propiedad

$\overline{O}$ es el punto de $\mathbb{R}^k$ que minimiza la suma de las distancias al cuadrado (error cuadrático medio, MSE), es decir, es la solución de

$$\min_{P\in\mathbb{R}^k} MSE = \sum_{i=}^{n} d^2(O_i, P),$$

donde d representa la distancia Euclídea, definida para dos vectores $x, y \in \mathbb{R}^k$ como

$$d(x,y) = \sqrt{\sum_{j=1}^{k}(x_j - y_j)^2}.$$

1.15) Estimador para la matriz de covarianzas V

Para estimar $\sigma_{i,j}$ usaremos

→ La covarianza muestral: $\hat{\sigma}_{i,j} = \frac{1}{n}\sum_{l=1}^{n}(X_{l,i} - \overline{X}_i)(X_{l,j} - \overline{X}_j)$

→ La cuasi-covarianza muestral:

$$\mathcal{S}_{i,j} = \frac{1}{n-1}\sum_{l=1}^{n}(X_{l,i} - \overline{X}_i)(X_{l,j} - \overline{X}_j)$$

Para estimar V usaremos:

→ $\hat{V} = (\hat{\sigma}_{i,j}) = \frac{1}{n}\sum_{l=1}^{n}(O_l - \overline{O})(O_l - \overline{O})'$

→ $\mathcal{S} = (\mathcal{S}_{i,j}) = \frac{1}{n-1}\sum_{n-1}^{n}(O_l - \overline{O})(O_l - \overline{O})'$

Se verifica que $E(\mathcal{S}) = V$ (estimador centrado de V).

1.15.1) Para una distribución normal

- Proposición

Si $X \longrightarrow \mathcal{N}_k(\mu, V)$ entonces se verifica que:

- $\overline{O} \longrightarrow \mathcal{N}_k\left(\mu, \frac{V}{n}\right)$
- $\overline{O}$ y $\hat{V}$ son los estimadores máximos verosímiles de μ y V, respectivamente.
- Además, $\overline{O}$ y $\hat{V}$ son independientes entre sí. Por tanto, también $\overline{O}$ y $\mathcal{S}$ son independientes entre sí.
- La distribución aleatoria

  $$n\hat{V} = (n-1)\mathcal{S}$$

  se conoce como distribuidor de Wishart.

- Test de normalidad multivariante: Test de Shapiro-Wilk

Para la aplicación de algunas técnicas multivariantes la hipótesis de normalidad es importante y debe ser contrastada.

$$H_0 : (X_1, \ldots, X_k) \to \mathcal{N}_k(\mu, V)$$
$$H_1 : (X_1, \ldots, X_k) \nrightarrow \mathcal{N}_k(\mu, V)$$

Podremos utilizar la función `mshapiro.test` de la librería `mvnormtest` de `R` para realizar el test de normalidad multivariante de Shapiro-Wilk.

→ Si aplicamos el test a los 50 datos simulados de la normal bivariante lógicamente obtnedremos un p-valor que apoya la hipótesis nula.

```
library("mvnormtest")
V <- matrix(c(1, 1/2,
              1/2, 1), nrow = 2, ncol = 2, byrow = TRUE)
mu <- c(0, 0)
seed = set.seed(2023)
d <- rmvnorm(50, mu, V)
mshapiro.test(t(d))
```

```
## [1]  0.6922
```

- Seguimos con **LifeCycleSavings**

Cálculo de las medias muestrales para cada variable.

```
mean(datos$sr); mean(datos$pop15); mean(datos$pop75);  mean(datos$dpi); mean(datos$
    ddpi)
```

```
## [1] 9.671
## [1] 35.0896
## [1] 2.293
## [1] 1106.758
## [1] 3.7576
```

O bien, podemos calcular todas las características de estas variables

```
summary(datos)
```

```
##        sr             pop15           pop75            dpi
##  Min.   : 0.600   Min.   :21.44   Min.   :0.560   Min.   :  88.94
##  1st Qu.: 6.970   1st Qu.:26.21   1st Qu.:1.125   1st Qu.: 288.21
##  Median :10.510   Median :32.58   Median :2.175   Median : 695.66
##  Mean   : 9.671   Mean   :35.09   Mean   :2.293   Mean   :1106.76
##  3rd Qu.:12.617   3rd Qu.:44.06   3rd Qu.:3.325   3rd Qu.:1795.62
##  Max.   :21.100   Max.   :47.64   Max.   :4.700   Max.   :4001.89
##       ddpi
##  Min.   : 0.220
##  1st Qu.: 2.002
##  Median : 3.000
##  Mean   : 3.758
##  3rd Qu.: 4.478
##  Max.   :16.710
```

Cálculo de la matriz de covarianzas muestrales

```
cov(d)
```

```
##           [,1]      [,2]
## [1,] 1.1101259 0.8347425
## [2,] 0.8347425 1.2075240
```

Cálculo de la matriz de correlaciones muestrales

En este caso es mejor usar correlaciones muestrales que eliminan el efecto de las unidades:

$$R_{i,j} = \frac{\mathcal{S}_{i,j}}{\mathcal{S}_i \mathcal{S}_j},$$

donde $\mathcal{S}_i = \sqrt{\mathcal{S}_{i,i}}$ y $\mathcal{S}_j = \sqrt{\mathcal{S}_{j,j}}$.

Cálculo de la matriz de correlaciones muestrales

```
1 cor(datos)
```

```
##                sr       pop15       pop75        dpi        ddpi
## sr      1.0000000 -0.45553809  0.31652112  0.2203589  0.30478716
## pop15 -0.4555381  1.00000000 -0.90847871 -0.7561881 -0.04782569
## pop75  0.3165211 -0.90847871  1.00000000  0.7869995  0.02532138
## dpi    0.2203589 -0.75618810  0.78699951  1.0000000 -0.12948552
## ddpi   0.3047872 -0.04782569  0.02532138 -0.1294855  1.00000000
```

Observamos que algunas variables tienen correlaciones positivas y otras negativas

RELACIÓN DE PROBLEMAS: VECTORES ALEATORIOS
ANÁLISIS ESTADÍSTICO MULTIVARIANTE
GRADO EN CIENCIA E INGENIERÍA DE DATOS

1. Sea (X, Y) un vector aleatorio con función de densidad conjunta

$$f(x,y) = \begin{cases} 1 & \text{si } 0 < x < 1, 0 < y < 1 \\ 0 & \text{en otro caso} \end{cases}$$

Hallar las distribuciones marginales y condicionadas.

2. Obtener las distribuciones marginales y condicionadas asociadas al vector aleatorio (X, Y) con función de densidad

$$f(x,y) = \begin{cases} 2 & \text{si } 0 < x < 1, 0 < y < x \\ 0 & \text{en otro caso} \end{cases}$$

3. Sea (X, Y) un vector aleatorio con función de densidad

$$f(x,y) = \begin{cases} \frac{3}{4}\left[xy + \frac{x^2}{2}\right] & \text{si } 0 < x < 1, 0 < y < 2 \\ 0 & \text{en otro caso} \end{cases}$$

Hallar la distribución marginal de X y la distribución de Y condicionada a $X = \frac{1}{2}$.

4. Sea $\mathbf{X} = (X_1, X_2)$ un vector aleatorio con función masa de probabilidad

$$P[X_1 = x_1, X_2 = x_2] = \frac{k}{2^{x_1+x_2}}, x_1, x_2 \in \mathbb{N},$$

donde k es una constante. Obtener las distribuciones marginales y condicionadas.

5. Calcular la función de densidad de una distribución normal bidimensional en (1, 1) si las medias son cero, las varianzas 1 y 4, y la covarianza 1.

6. Sea (X, Y) un vector aleatorio con distribución uniforme en el cuadrado unidad, $[0, 1] \times [0, 1]$, con función de densidad conjunta

$$f(x,y) = \begin{cases} 1 & \text{si } 0 < x < 1, 0 < y < 1 \\ 0 & \text{en otro caso} \end{cases}$$

Calcular el valor esperado de $g(X, Y) = XY^2$, es decir, $E[XY^2]$.

7. (X, Y) vector aleatorio discreto con función masa de probabilidad conjunta:

$X \backslash Y$	1	2
1	1/9	2/9
2	2/9	4/9

   a) Calcular $E[X + Y]$, $E[2X + 3Y]$.
   b) Obtener el vector de medias, la matriz de covarianzas y la matriz de correlaciones del vector (X, Y).
   c) ¿Son independientes? ¿Están incorreladas?

8. Demostrar que el vector de medias muestral es el punto de $\mathbb{R}^k$ que minimiza la suma de las distancias al cuadrado (error cuadrático medio, MSE).

1) Sea (X,Y) un vector aleatorio con función de densidad conjunta

$$f(x,y)=\begin{cases}1 & \text{si } 0<x<1,\ 0<y<1\\ 0 & \text{en otro caso}\end{cases}$$

Hallar las distribuciones marginales y condicionadas

$$\underset{0<x<1}{f_X(x)}=\int_{\infty}^{+\infty} f(x,y)\,\mathrm{d}y=\int_{-\infty}^{0} 0\,\mathrm{d}y+\int_0^1 1\,\mathrm{d}y+\int_0^{+\infty} 0\,\mathrm{d}y=[y]_{y=0}^{y=1}=1 \qquad f_X(x)=\begin{cases}1 & \text{si } 0<x<1\\ 0 & \text{en otro caso}\end{cases}$$

$$\underset{0<y<1}{f_Y(y)}=\int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}x=\int_0^1 1\,\mathrm{d}x=[x]_{x=0}^{x=1}=1 \qquad f_Y(y)=\begin{cases}1 & \text{si } 0<y<1\\ 0 & \text{en otro caso}\end{cases}$$

$$f(x,y)=\begin{cases}1 & 0<x<1,\ 0<y<1\\ 0 & \text{en otro caso}\end{cases}$$

$$\underset{f_X(x^*)>0}{y|x=x^*}\longrightarrow \underset{0<x^*<1}{f_{y|x=x^*}}=\frac{f(x^*,y)}{f_X(x^*)}=\begin{cases}1 & 0<y<1\\ 0 & \text{en otro caso}\end{cases}$$

$f(x,y)=f_X(x)\cdot f_Y(y)$ $\quad X$ e Y independientes

2) Obtener las distribuciones marginales y condicionadas asociadas al vector aleatorio (X,Y) con función de densidad

$$f(x,y)=\begin{cases}2 & \text{si } 0<x<1,\ 0<y<x\\ 0 & \text{en otro caso}\end{cases}$$

$$\underset{0<x<1}{f_X(x)}=\int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}y=\int_0^x 2\,\mathrm{d}y=[2y]_{y=0}^{y=x}=2x\longrightarrow\begin{cases}2x & \text{si } 0<x<1\\ 0 & \text{en otro caso}\end{cases}$$

$$\underset{0<y<1}{f_Y(y)}=\int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}x=\int_y^1 2\,\mathrm{d}x=[2x]_{x=y}^{x=1}=2-2y\longrightarrow\begin{cases}2-2y & \text{si } 0<y<1\\ 0 & \text{en otro caso}\end{cases}$$

Los recintos son dependientes.

$y|x=x^*$

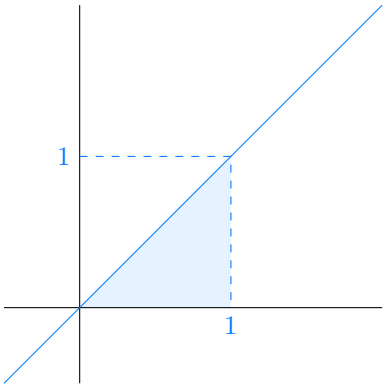
$f_X(x^*)>0$

$$\underset{0<x^*<1}{f_{y|x=x^*}}(y|x^*)=\frac{f(x^*,y)}{f(x^*)}=\begin{cases}\dfrac{2}{2x^*} & 0<y<x^*\\ 0 & \text{en otro caso}\end{cases}=\begin{cases}\dfrac{1}{x^*} & 0<y<x^*\\ 0 & \text{en otro caso}\end{cases}$$

$x|y=y^*$

$f_Y(y^*)>0$

$$\underset{0<y^*<1}{f_{x|y=y^*}}(x|y^*)=\frac{f(x,y^*)}{f_Y(y^*)}=\begin{cases}\dfrac{2}{2-2y^*} & \text{si } y^*<x<1\\ 0 & \text{en otro caso}\end{cases}=\begin{cases}\dfrac{1}{1-y^*} & \text{si } y^*<x<1\\ 0 & \text{en otro caso}\end{cases}$$



3) Sea (X, Y) un vector aleatorio con función de densidad

$$f(x,y) = \begin{cases} \frac{3}{4}\left[xy + \frac{x^2}{2}\right] & \text{si } 0 < x < 1,\ 0 < y < 2 \\ 0 & \text{en otro caso} \end{cases}$$

Hallar la distribución marginal de X y la distribución de Y condicionada a $X = \frac{1}{2}$.

$$f_{\substack{X \\ 0<x<1}}(x) = \int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}y = \int_0^2 \frac{3}{4}\left[xy + \frac{x^2}{2}\right] = \frac{3}{4}\left[\frac{xy^2}{2} + \frac{x^2}{2}\cdot y\right]_{y=0}^{y=2} = \frac{3}{4}\left(2x + x^2\right) \longrightarrow \begin{cases} \frac{3}{4}\left(2x + x^2\right) & \text{si } 0 < x < 1 \\ 0 & \text{en otro caso} \end{cases}$$

$y|x = x^*$

$$f_{\substack{y|x=x^* \\ 0<x^*<1}}(y|x^*) = \frac{f(x^*,y)}{f(x^*)} = \frac{\frac{3}{4}\left(x^*y + \frac{(x^*)^2}{2}\right)}{\frac{3}{4}\left(2x^* + (x^*)^2\right)} = \frac{x^*y + \frac{(x^*)^2}{2}}{2x^* + \frac{(x^*)^2}{2}} = \frac{x^*y + (x^*)^2}{4x^* + 2(x^*)^2} \xrightarrow{x^*=\frac{1}{2}} \frac{\frac{1}{2}y + \frac{1}{8}}{2\cdot\frac{1}{2} + \frac{1}{4}} = 2\cdot\frac{y + \frac{1}{4}}{5}$$

$$\begin{cases} 2\cdot\frac{y+\frac{1}{4}}{5} & \text{si } 0 < y < 2 \\ 0 & \text{en otro caso} \end{cases}$$

4) Sea $X = (X_1, X_2)$ un vector aleatorio con función masa de probabilidad

$$P[X_1 = x_1, X_2 = x_2] = \frac{k}{2^{x_1+x_2}}, x_1, x_2 \in \mathbb{N}$$

donde k es una constante. Obtener las distribuciones marginales y condicionadas.

$P[X_1 = x_1, X_2 = x_2] = \frac{k}{2^{x_1+x_2}}$, $x_1, x_2 \in \mathbb{N}$ (incluido el 0)

$$P[\underset{x_1\in\mathbb{N}}{X_1 = x_1}] = \sum_{x_2\in\mathbb{N}} \frac{k}{2^{x_1+x_2}} = \frac{k}{2^{x_1}} \sum_{x_2\in\mathbb{N}} \frac{1}{2^{x_2}} = \frac{k}{2^{x_1}} \cdot \frac{1}{1-\frac{1}{2}} = \frac{2k}{2^{x_1}}$$

$$P[\underset{x_1^*\in\mathbb{N}}{X_2 = x_2}|X_1 = x_1^*] = \frac{P[X_1 = x_1^*, X_2 = x_2]}{P[X_1 = x_1]} = \begin{cases} \frac{\frac{k}{2^{x_1+x_2}}}{\frac{2k}{2^{x_1}}} = \frac{1}{2\cdot 2^{x_2}} & x_2 \in \mathbb{N} \\ 0 & \text{en otro caso} \end{cases}$$

5) Calcular la función de densidad de una distribución normal bidimensional en $(1, 1)$ si las medias son cero, las varianzas 1 y 4, y la covarianza 1.

Fórmula de la función de densidad de una distribución normal bidimensional:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right)$$

$$\mu_x = \mu_y = 0$$
$$\sigma_x^2 = 1$$
$$\sigma_y^2 = 4$$
$$\rho = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{1}{\sqrt{1}\cdot\sqrt{4}} = \frac{1}{2}$$

$$\begin{aligned} f(1,1) &= \frac{1}{2\pi \cdot 1 \cdot 2\sqrt{1-\left(\frac{1}{2}\right)^2}} \exp\left(-\frac{1}{2\left(1-\left(\frac{1}{2}\right)^2\right)} \cdot \left[1^2 + \frac{1^2}{4} - \frac{2\cdot\frac{1}{2}}{1\cdot 2}\right]\right) \\ &= \frac{1}{2\pi\sqrt{3}} \exp\left(-\frac{2}{3}\cdot\frac{3}{4}\right) \\ &= \frac{1}{2\pi\sqrt{3}} \exp\left(-\frac{1}{2}\right) \simeq \boxed{0.0557} \end{aligned}$$

$$\underset{x\in\mathbb{R}^k}{f(x)} = \frac{1}{|V|(2\pi)^k} e^{-\frac{1}{2}(x-\mu)^\intercal \Sigma^{-1}(x-\mu)}$$

$$V = \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix} \qquad |V| = 3$$

$$\text{Adj}(V^\intercal) = \begin{pmatrix} 4 & -1 \\ -1 & 1 \end{pmatrix} \qquad V^{-1} = \frac{1}{|V|}\text{Adj}(V^\intercal) = \begin{pmatrix} \frac{4}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

$$\begin{pmatrix} x-0 & y-0 \end{pmatrix} \cdot \begin{pmatrix} \frac{4}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{1}{3} \end{pmatrix} \cdot \begin{pmatrix} x-0 \\ y-0 \end{pmatrix} = \begin{pmatrix} \frac{4}{3}x\frac{y}{3} & -\frac{x}{3}+\frac{y}{3} \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} = \frac{4}{3}x^2 - \frac{xy}{3} - \frac{x}{y} + \frac{y^2}{3}$$

$$f(x,y) = \frac{1}{\sqrt{3(2\pi)^k}} \cdot e^{-\frac{1}{2}\left(\frac{4}{3}x^2 - \frac{2xy}{3} - \frac{x}{y} + \frac{y^2}{3}\right)} \longrightarrow f(1,1) \simeq 0.0557$$

6) Sea (X,Y) un vector aleatorio con distribución uniforme en el cuadradado unidad, $[0,1]\times[0,1]$, con función de densidad conjunta

$$f(x,y) = \begin{cases} 1 & \text{si } 0 < x < 1,\ 0 < y < 1 \\ 0 & \text{en otro caso} \end{cases}$$

Calcular el valor esperado de $g(X,Y) = XY^2$, es decir, $E[XY^2]$.

El valor esperado de una función $g(X,Y)$ para una variable aleatoria conjunta (X,Y) se define como:

$$E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

En este caso, $g(X,Y) = XY^2$ y la función de densidad conjunta $f(x,y)$ es 1 para $0 < x < 1$ y $0 < y < 1$, y 0 en otro caso. Por lo tanto, el valor esperado se convierte en:

$$E[XY^2] = \int_0^1\int_0^1 xy^2\,\mathrm{d}x\,\mathrm{d}y$$

Resolviendo la integral obtenemos:

$$E[XY^2] = \int_0^1 \left[\frac{1}{2}x^2y^2\right]_0^1 \mathrm{d}y = \int_0^1 \frac{1}{2}y^2\,\mathrm{d}y = \left[\frac{1}{6}y^3\right]_0^1 = \frac{1}{6}$$

Por lo tanto, el valor esperado de XY^2 es $\frac{1}{6}$.

$$E[XY^2] = E[X]\cdot E[Y^2] = (*) = \frac{1}{2}\cdot\frac{1}{3} = \frac{1}{6}$$
$$E[X] = \int_0^1 x\,\mathrm{d}x = \left[\frac{x^2}{2}\right]_0^1 = \frac{1}{2}$$
$$E[Y^2] = \int_0^1 y^2\,\mathrm{d}y = \left[\frac{y^3}{3}\right]_0^1 = \frac{1}{3}$$

7) (X,Y) vector aleatorio discreto con función masa de probabilidad conjunta:

X \ Y	1	2
1	$\frac{1}{9}$	$\frac{2}{9}$
2	$\frac{2}{9}$	$\frac{4}{9}$

a) Calcular $E[X+Y], E[2X+3Y]$.

$$E[X+Y] = E[X] + E[Y] = \frac{5}{3} + \frac{5}{3} = \frac{10}{3}$$
$$E[X] = 1 \cdot P(X=1) + 2 \cdot P(X=2) = 1 \cdot \left(\frac{1}{9} + \frac{2}{9}\right) + 2 \cdot \left(\frac{2}{9} + \frac{4}{9}\right) = \frac{1}{3} + \frac{4}{3} = \frac{5}{3}$$
$$E[Y] = 1 \cdot P(Y=1) + 2 \cdot P(Y=2) = 1 \cdot \left(\frac{1}{9} + \frac{2}{9}\right) + 2 \cdot \left(\frac{2}{9} + \frac{4}{9}\right) = \frac{1}{3} + \frac{4}{3} = \frac{5}{3}$$

$$E[2X+3Y] = 2E[X] + 3E[Y] = 2 \cdot \frac{5}{3} + 3 \cdot \frac{5}{3} = \frac{25}{3}$$

b) Obtener el vector de medias, la matriz de covarianzas y la matriz de correlaciones del vector (X, Y).

- Vector de medias:

$$\mu = \begin{bmatrix} E[X] \\ E[Y] \end{bmatrix} = \begin{bmatrix} \frac{5}{3} \\ \frac{5}{3} \end{bmatrix} = \frac{5}{3} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

- Matriz de covarianzas:

$$\Sigma = \begin{bmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(Y,X) & \mathrm{Var}(Y) \end{bmatrix} = \begin{bmatrix} \frac{2}{9} & 0 \\ 0 & \frac{2}{9} \end{bmatrix}$$

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2 = 3 - \left(\frac{5}{3}\right)^2 = \frac{2}{9}$$
$$\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y] = \frac{25}{9} - \frac{5}{3} \cdot \frac{5}{3} = 0$$
$$\mathrm{Var}(Y) = E[Y^2] - (E[Y])^2 = 3 - \left(\frac{5}{3}\right)^2 = \frac{2}{9}$$
$$E[X^2] = 1^2 \cdot P(X=1) + 2^2 \cdot P(X=2) = 1^2 \cdot \left(\frac{1}{9} + \frac{2}{9}\right) + 2^2 \cdot \left(\frac{2}{9} + \frac{4}{9}\right) = \frac{1}{3} + \frac{8}{3} = 3$$
$$E[Y^2] = 1^2 \cdot P(Y=1) + 2^2 \cdot P(Y=2) = 1^2 \cdot \left(\frac{1}{9} + \frac{2}{9}\right) + 2^2 \cdot \left(\frac{2}{9} + \frac{4}{9}\right) = \frac{1}{3} + \frac{8}{3} = 3$$

$$E[XY] = 1 \cdot 1 \cdot P(X=1, Y=1) + 1 \cdot 2 \cdot P(X=1, Y=2) + 2 \cdot 1 \cdot P(X=2, Y=1) + 2 \cdot 2 \cdot P(X=2, Y=2) = \frac{1}{9} + 2 \cdot \frac{2}{9} + 2 \cdot \frac{2}{9} + 4 \cdot \frac{4}{9} = \frac{25}{9}$$

- Matriz de correlaciones:

$$R = \begin{bmatrix} 1 & \mathrm{Corr}(X,Y) \\ \mathrm{Corr}(Y,X) & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I$$

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)}} = \frac{0}{\sqrt{\frac{2}{9} \cdot \frac{2}{9}}} = 0$$

c) ¿Son independientes? ¿Están incorreladas?

Las variables aleatorias X e Y serán independientes si se cumple la condición: $P(X=x, Y=y) = P(X=x) \cdot P(Y=y) \forall x, y \in \mathbb{N}$

$$P(X=1, Y=1) = P(X=1) \cdot P(Y=1) \longrightarrow \frac{1}{9} = \left(\frac{1}{9} + \frac{2}{9}\right) \cdot \left(\frac{1}{9} + \frac{2}{9}\right) \longrightarrow \frac{1}{9} = \frac{1}{3} \cdot \frac{1}{3}$$

Por lo tanto, son independientes.

Las variables aleatorias X e Y están incorreladas si su covarianza vale 0. En este caso sí están incorreladas.

8) Demostrar que el vector de medias muestral es el punto $\mathbb{R}^k$ que minimiza la suma de las distancias al cuadrado (error cuadrático medio, MSE).

Tema 2: Regresión lineal simple y múltiple

- Introducción

Objetivo: predecir una variable numérica a partir de k variables numéricas (variables predictoras) minimizando el error en la predicción.

Para ello necesitamos disponer de una muestra en la que se conozcan dichas variables (aprendizaje supervisado), esta muestra se usará para elegir el mejor modelo y para validar su fiabilidad.

- Planteamiento

Se trata de predecir el valor (numérico) de una variable aleatoria (v.a.) Y a partir de unas variables predictoras $X_1, \ldots, X_k$.

Para ello usaremos una función predictora lineal

$$h_\theta(x_1, \ldots, x_k) := \theta_0 + \theta_1 x_1 + \cdots + \theta_k x_k,$$

donde $\theta = (\theta_0, \ldots, \theta_k)' \in \mathbb{R}^{k+1}$ serán los parámetros del modelo que se deben elegir de forma que la estimación de Y sea óptima (el error sea mínimo).

→ Una sola variable predictora ⟶ Regresión lineal simple.

→ Más de una variable predictora ⟶ Regresión lineal múltiple.

2.1) Estimación de los parámetros

2.1.1) Muestra

Para ello necesitamos disponer de una muestra (training sample) de esas $k+1$ variables sobre n individuos.

Una realización de la muestra se denotará como

$$\left(x_1^{(i)}, \ldots, x_k^{(i)}, y^{(i)}\right), \quad i = 1, \ldots, n,$$

que equivale a la notación introducida en el tema anterior

$$(x_{i,1}, \ldots, x_{i,k}, y_i), \quad i = 1, \ldots, n.$$

Los datos se representan en forma de tabla:

i	x_1	x_2	$\cdots$	x_k	y
$\mathbf{o}_1$	$x_1^{(1)}$	$x_2^{(1)}$	$\cdots$	$x_k^{(1)}$	y_1
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\mathbf{o}_i$	$x_1^{(i)}$	$x_2^{(i)}$	$\cdots$	$x_k^{(i)}$	y_i
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\mathbf{o}_n$	$x_1^{(n)}$	$x_2^{(n)}$	$\cdots$	$x_k^{(n)}$	y_n

i	x_1	x_2	$\cdots$	x_k	y
$\mathbf{o}_1$	$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,k}$	y_1
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\mathbf{o}_i$	$x_{i,1}$	$x_{i,2}$	$\cdots$	$x_{i,k}$	y_i
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\mathbf{o}_n$	$x_{n,1}$	$x_{n,2}$	$\cdots$	$x_{n,k}$	y_n

2.2) Conjunto de datos `USArrests`

Cargamos los datos

Podemos visualizar los datos con `view(d)` y las primeras filas con `head(d)`.

```r
d <- USArrests
head(d, n = 6)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

¿Qué información está recogida en el conjunto de datos?

Con la instrucción `help(USArrests)` conocemos qué información está contenida en el conjunto:

- `Murder:` Ratios de arrestos por asesinatos por cada 100 000 residentes en cada uno de los 50 estados de la unión.
- `Assault:` Ratios de arrestos por agresión por cada 100 000 residentes en cada uno de los 50 estados de la unión.
- `UrbanPop:` Porcentaje de población que vive en áreas urbanas.
- `Rape:` Ratios de arrestos por violación por cada 100 000 residentes en cada uno de los 50 estados de la unión.

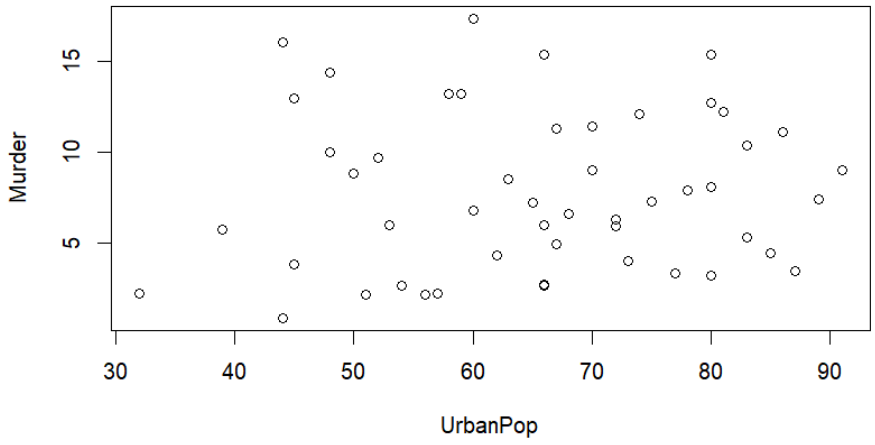
2.2.1) Objetivo

Predecir el ratio de arrestos por asesinatos (Y =`Murder`) en función de la variable X =`UrbanPop`.

Para visualizar la relación entre estas variables podemos representarlas situando X en el eje horizontal e Y en el vertical.

```r
x <- d$UrbanPop #Elegimos x
y <- d$Murder #Elegimos y
plot(x, y, xlab = 'UrbanPop', ylab = 'Murder')
```

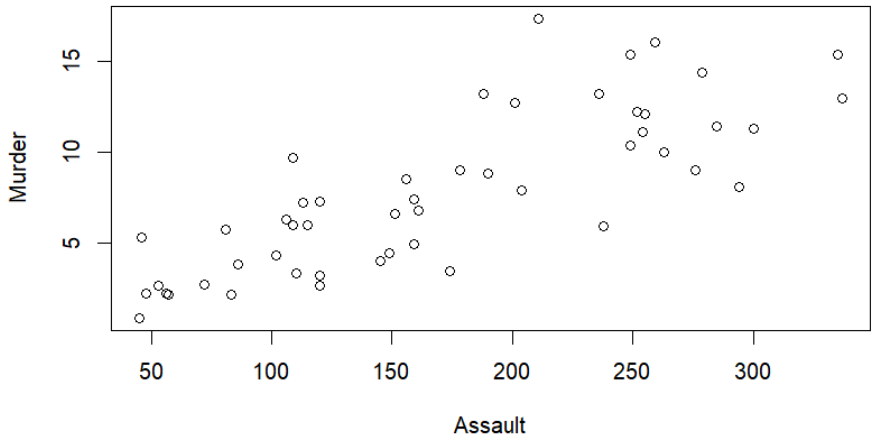
Podemos observar que no parece existir ninguna relación entre las variables `Murder` y `UrbanPop` por lo que la predicción no será muy buena.



Si usamos como predictor la variable `Assault` y representamos gráficamente:

Ahora sí se aprecia una relación lineal (creciente entre ambas variables.)

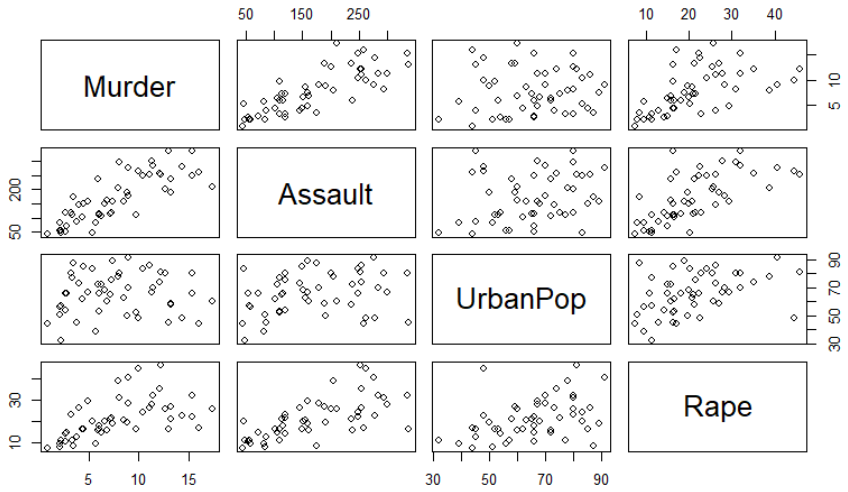
```
x <- d$Assault #Elegimos x
y <- d$Murder #Elegimos y
plot(x, y, xlab = 'Assault', ylab = 'Murder')
```



2.2.2) Matriz de gráficos

Podemos representar conjuntamente todas las variables con gráficos bidimensionales para cada pareja de variables.

```
plot(d)
```



2.2.3) Resumen numérico de las variables

Podemos obtener los estadísticos descriptivos de estas variables con `summary(d)` que incluyen los extremos (mínimo y máximo), los cuartiles, la mediana y la media.

```
summary(d)
```

```
##      Murder          Assault         UrbanPop          Rape
##  Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30
##  1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07
##  Median : 7.250   Median :159.0   Median :66.00   Median :20.10
##  Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23
##  3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18
##  Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00
```

Siempre es buena idea usar medidas descriptivas y gráficas para analizar los datos antes de aplicar un procedimiento estadístico multivariante.

2.2.4) Estimación de las varianzas y covarianzas

Para calcular una estimación de las varianzas y covarianzas utilizaremos la función `var` (`R` obtiene las cuasivarianzas).

```r
#calculo directo de la varianza y cuasivarianza para Murder
mu <- mean(d$Murder)
n <- length(d)
hat_sigma = sum((d$Murder-mu)^ 2)/n #varianza muestral
S = sum((d$Murder-mu)^ 2)/(n-1) #cuasivarianza muestral
#calculo de la matriz de covarianzas
var(d)
```

```
##              Murder   Assault   UrbanPop      Rape
## Murder    18.970465  291.0624   4.386204  22.99141
## Assault  291.062367 6945.1657 312.275102 519.26906
## UrbanPop   4.386204  312.2751 209.518776  55.76808
## Rape      22.991412  519.2691  55.768082  87.72916
```

2.2.5) Modelo completo

Podemos incluir todas las variables en el modelo considerado

$$h_\theta = \theta_0 + \theta_1 \text{Assault} + \theta_2 \text{UrbanPop} + \theta_3 \text{Rape}$$

donde $\theta = (\theta_0, \theta_1, \theta_2, \theta_3)' \in \mathbb{R}^4$ son los parámetros del modelo que debemos ajustar para obtener las mejores aproximaciones posibles.

Los casos en los que solo usamos una variable están incluidos en este modelo haciendo que los parámetros de las otras variables sean cero.

También podemos intentar mejorar estas aproximaciones considerando otras funciones h (no lineales).

2.3) Regresión lineal simple

2.3.1) Modelo teórico

Partiremos de un vector aleatorio (X, Y).

Objetivo: Construir una nueva variable $h(X)$ que se parezca (aproxime) a Y.

Los errores (residuos) serán otra variable aleatoria

$$R = Y - h(X)$$

(notemos que pueden ser positivos o negativos).

Existen diversas reglas para determinar una función objetivo que mida cómo son esos errores y trate de minimizarlos.

La más usada es el denominado error cuadrático medio (EMC) definido como:

$$EMC = E\left[(h(X) - Y)^2\right]$$

(MSE, Mean Square Error)

2.3.2) Función óptima

Supongamos que (X, Y) tiene una distribución absolutamente continua con función de densidad conjunta f y marginales f_X y f_Y.

Entonces se puede demostrar que la función h que minimiza el EMC es

$$h_{\text{opt}}(x) = E(Y|X = x) = \int_{-\infty}^{+\infty} y f_{Y|X}(y|x)\, \mathrm{d}y,$$

donde

$$f_{Y|X}(y|x) = f(x, y)|f_X(x),$$

para tales $f_X(x) > 0$, es la función de densidad condicionada de $(Y|X = x)$.

Esta función se denomina curva de regresión y es el mejor predictor de Y dado X según el ECM.

2.4) Caso de normalidad

El vector (X, Y) tiene una distribución normal $\mathcal{N}_2(\mu, V)$:

$\mu = (\mu_1, \mu_2)'$ es el vector de medias (A' representa la traspuesta de la matriz A), donde

$$\mu_1 = \mu_X = E[X]$$
$$\mu_2 = \mu_Y = [Y]$$

$V = \begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_{2,2} \end{pmatrix}$ es la matriz de varianzas-covarianzas, donde

$$\sigma_{1,1} = \sigma_X^2 = \mathrm{Var}(X) = E[(X - \mu_X)^2],$$
$$\sigma_{2,2} = \sigma_Y^2 = \mathrm{Var}(Y) = E[(Y - \mu_Y)^2],$$
$$\sigma_{1,2} = \sigma_{2,1} = \sigma_{X,Y} = \mathrm{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)].$$

Entonces la distribución condicionada $(Y|X = x)$ se comporta también como una distribución normal,

$$(Y|X = x) \longrightarrow \mathcal{N}_1(\overline{\mu}, \overline{\sigma}^2),$$

con

$$h_{\text{opt}}(x) = \overline{\mu} = E(Y|X=x) = \mu_2 + \frac{\sigma_{1,2}}{\sigma_{1,1}}(x-\mu_1)$$
$$= \mu_Y + \frac{\text{Cov}(X,Y)}{\sigma_X^2}(x-\mu_X)$$
$$\overline{\sigma}^2 = \text{Var}(Y|X=x) = \sigma_{2,2} - \frac{\sigma_{1,2}^2}{\sigma_{1,1}} = \sigma_Y^2 - \frac{\text{Cov}(X,Y)^2}{\sigma_X^2}.$$

- Observaciones

Bajo la hipótesis de normalidad, la curva de regresión h_{opt} es siempre una recta y la varianza $\overline{\sigma}^2$ no depende de x.

Los residuos condicionados $R_x = R|X=x$ también serán normales $R_x \longrightarrow \mathcal{N}(0,\overline{\sigma}^2)$ e idénticamente distribuidos.

La curva (recta) de regresión para predecir Y en función de X se puede escribir como

$$\frac{y-\mu_Y}{\sigma_Y} = \rho_{X,Y}\frac{x-\mu_X}{\sigma_X},$$

donde $\rho_{X,Y} = \dfrac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y}$ es el coeficiente de correlación lineal de Pearson.

La recta siempre pasa por el punto (μ_X, μ_Y).

2.4.1) Coeficiente de correlación de Pearson

- Propiedades

El signo de la pendiente de la recta de regresión siempre coincide con el signo de $\rho_{X,Y}$.

Se verifica que:

$$\overline{\sigma}^2 = \sigma_Y^2 - \frac{\text{Cov}(X,Y)^2}{\sigma_X^2\sigma_Y^2}\sigma_Y^2 = (1-\rho_{X,Y}^2)\sigma_Y^2 \geq 0,$$

por lo que $-1 \leq \rho_{X,Y} \leq 1$.

Cuando $\rho_{X,Y} = \pm 1$ tendremos ajustes perfectos con residuos nulos.

La recta (curva) para predecir X a partir de Y se calcula de forma similar y no coincide con la curva para predecir Y a partir de X que acabamos de calcular salvo cuando $\rho_{X,Y} = \pm 1$.

2.5) Caso de no normalidad

- Observaciones

Para otras distribuciones bivariantes la curva de regresión no tiene por qué ser una recta.
Cuando X e Y sean independientes, Y e $(Y|X=x)$ tienen la misma distribución y la curva óptima

$$h_{\text{opt}}(x) = E(Y|X=x) = E(Y)$$

es constante (por lo que también es una recta).

→ En este caso el valor de X no influye en la predicción sobre Y.

→ $\rho_{X,Y} = 0$ (recta horizontal).

2.6) Restricción sobre la función h

En **Regresión Lineal Simple** supondremos que la función h es una recta

- Limitamos nuestra función h a una recta, es decir,

$$h_\theta(x) \coloneqq \theta_0 + \theta_1 x$$

- Usamos como criterio minimizar el ECM.
- El objetivo será

$$\min_{\theta_0,\theta_1} J(\theta_0, \theta_1),$$

donde

$$J(\theta_0, \theta_1) \coloneqq E[(h_\theta(X) - Y)^2] = E[(\theta_0 + \theta_1 X + Y)^2]$$

se conoce como función costo y $J(\theta_0, \theta_1) \geq 0$.

→ Por lo tanto, se trata de minimizar una función costo

2.7) Minimizar la función costo

2.7.1) Ecuaciones normales de la recta

La función $J(\theta)$ es convexa por lo que tendrá un único mínimo que se puede obtener resolviendo el sistema

$$\begin{aligned} \frac{\partial J(\theta_0, \theta_1)}{\partial \theta_0} &= 2E[\theta_0 + \theta_1 X - Y] = 0 \\ \frac{\partial J(\theta_0, \theta_1)}{\partial \theta_1} &= 2E[(\theta_0 + \theta_1 X - Y)X] = 0 \end{aligned}$$

Estas ecuaciones se conocen como ecuaciones normales.

De la primera ecuación obtenemos

$$\theta_0 = E(Y) - \theta_1 E(X)$$

(con lo que la recta pasará por el punto formado con las medias).

De la segunda

$$\theta_0 E(X) + \theta_1 E(X^2) = E(XY)$$

Y sustituyendo la primera en la segunda, se obtiene

$$E(X)E(Y) - \theta_1 E^2(X) + \theta_1 E(X^2) = E(XY),$$

es decir,

$$\theta_1 \operatorname{Var}(X) = \operatorname{Cov}(X, Y),$$

puesto que $\operatorname{Var}(X) = E(X^2) - E^2(X)$ y $\operatorname{Cov}(X, Y) = E(XY) - E(X)E(Y)$.

Con lo cual,

→ $\hat{\theta}_1 = \dfrac{\operatorname{Cov}(X, Y)}{\operatorname{Var}(X)} = \dfrac{\sigma_{X,Y}}{\sigma_X^2}$

→ $\hat{\theta}_0 = E(Y) - \hat{\theta}_1 E(X) = E(Y) - E(X)\dfrac{\operatorname{Cov}(X, Y)}{\operatorname{Var}(X)} = \mu_Y - \mu_X \dfrac{\sigma_{X,Y}}{\sigma_X^2}$

2.8) Expresión de la recta

2.8.1) Recta de regresión para predecir Y en función de X

En el punto $(\hat{\theta}_0, \hat{\theta}_1)$ se alcanza el mínimo de la función J,

$$\min_{\theta_0,\theta_1} J(\theta_0, \theta_1) = J(\hat{\theta}_0, \hat{\theta}_1)$$

Expresión de la recta:

$$h_{\hat{\theta}}(x) = \mu_Y - \mu_X \frac{\sigma_{X,Y}}{\sigma_X^2} + \frac{\sigma_{X,Y}}{\sigma_X^2} x = \mu_Y + \frac{\sigma_{X,Y}}{\sigma_X^2}(x - \mu_X).$$

(Note que la fórmula es la misma que la de la curva de regresión de la normal).

Otra expresión:

$$\frac{y - \mu_Y}{\sigma_Y} = \rho_{X,Y} \frac{x - \mu_X}{\sigma_X}$$

donde $\rho_{X,Y} = \dfrac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$ es el coeficiente de correlación lineal de Pearson.

La variable aleatoria

$$\hat{Y} = h_{\hat{\theta}}(X) = \hat{\theta}_0 + \hat{\theta}_1 X$$

se usará para estimar Y.

Los residuos se definen como $R = Y - \hat{Y}$.

Se verifica:

→ $E(\hat{Y}) = E(Y)$

→ $E(R) = 0$

2.9) Descomposición de la varianza

2.9.1) Relaciones entre las varianzas

Expresando $Y = \hat{Y} + R$, se tiene que:

$$\sigma_y = \mathrm{Var}(Y) = \mathrm{Var}(\hat{Y}) + \mathrm{Var}(R)$$

Puesto que

$$\mathrm{Var}(\hat{Y}) = \hat{\theta}_1^2 \sigma_X^2 = \frac{\sigma_{X,Y} P2}{\sigma_X^2} = \rho_{X,Y}^2 \sigma_Y^2$$
$$\mathrm{Var}(R) = \mathrm{Var}(Y - \hat{\theta}_1 X) = (1 - \rho_{X,Y}^2)\sigma_Y^2.$$

Es decir, la información (varianza) contenida en Y se descompone como

$$\sigma_Y^2 = \rho_{X,Y}^2 \sigma_Y^2 + (1 - \rho_{X,Y}^2)\sigma_Y^2.$$

2.10) Coeficiente de determinación

- Definición

El coeficiente de determinación $d_{X,Y} = \rho_{X,Y}^2$ es el porcentaje (en tanto por 1) de la información de Y explicada por la recta de regresión (por relaciones lineales de X). Denotado habitualmente en los paquetes estadísticos por R^2.

Análogamente, $1 - d_{X,Y} = 1 - \rho_{X,Y}^2$ indicaría la parte de Y no explicada por esa recta y que se queda en el residuo.

Además, se tiene que

$$E(\hat{Y} R) = 0,$$

es decir, la variable que se obtiene con la recta de regresión y los residuos son incorrelados.

Bajo normalidad, ambas variables serán normales (por ser combinaciones lineales de X e Y) y, por lo tanto, serán independientes.

2.11) Inferencia y predicción

- Una muestra

En la práctica tanto la distribución conjunta (densidad) de (X, Y) como todas esas medidas serán desconocidas por lo que tendrán que ser estimadas a partir de una muestra de esas variables (training sample).

Si la muestra es grande, podemos extraer algunos datos (no usados en el cálculo de la recta) para comprobar cómo de fiables serán nuestras estimaciones.

La muestra se denotará como

$$\left(x^{(i)}, y^{(i)}\right), \quad i = 1, \ldots, n,$$

donde n será el tamaño muestral.

Los datos de cada variable se representarán como columnas y todos lo datos como una matriz D.

2.12) Función costo empírica

2.12.1) Objetivo

Queremos aproximar los valores de Y mediante una recta (función lineal) de X, es decir,

$$h_\theta \coloneqq \theta_0 + \theta_1 x,$$

donde $\theta = (\theta_0, \theta_1)$ son parámetros desconocidos.

Para calcular estos parámetros minimizaremos una función coste empírica J que nos mida el error cometido.

La más utilizada es el error cuadrático medio (o una función proporcional a él), por ejemplo podemos considerar

$$J(\theta_0, \theta_1) \coloneqq \frac{1}{2n} \sum_{i=1}^{n} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 = \frac{1}{2n} \sum_{i=1}^{n} \left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right)^2$$

El objetivo es minimizar esta función en $\mathbb{R}^2$.

2.12.2) Diferenciar J

Para obtener la solución exacta debemos diferencias J con respecto a los parámetros obteniendo

$$\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \frac{1}{n} \sum_{i=1}^{n} \left(h\theta\left(x^{(i)}\right) - y^{(i)}\right)$$
$$\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \frac{1}{n} \sum_{i=1}^{n} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x^{(i)}$$

Igualando a cero obtenemos las ecuaciones normales empíricas:

$$\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \frac{1}{n} \sum_{i=1}^{n} \left(h\theta\left(x^{(i)}\right) - y^{(i)}\right) = \frac{1}{n} \sum_{i=1}^{n} \left(\theta_0 + \theta x^{(i)} - y^{(i)}\right) \cdot 1 = 0$$
$$\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \frac{1}{n} \sum_{i=1}^{n} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x^{(i)} = \frac{1}{n} \sum_{i=1}^{n} \left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right) x^{(i)} = 0$$

2.12.3) Solución exacta

De la primera ecuación,

$$\theta_0 + \theta_1 \overline{x} - \overline{y} = 0$$

donde $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$ e $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y^{(i)}$ son las medias muestrales de x e y, respectivamente.

La solución óptima pasa por el punto medio $(\overline{x}, \overline{y})$ (individuo promedio).

De la segunda,

$$\theta_0 \overline{x} + \theta_1 a(x,x) - a(x,y) = 0,$$

donde $a(x,x) = \frac{1}{n}\sum_{i=1}^{n} \left(x^{(i)}\right)^2$ y $a(x,y) = \frac{1}{n}\sum_{i=1}^{n} x^{(i)} y^{(i)}$.

Resolviendo este sistema de ecuaciones obtenemos

$$\theta_1 \left(a(x,x) - (\overline{x})^2\right) = a(x,y) - \overline{x}\overline{y}$$

Obtenemos:

$$\hat{\theta}_1 = \frac{s_{x,y}}{s_x^2}, \quad \hat{\theta}_0 = \overline{y} - \hat{\theta}_1 \overline{x},$$

donde $s_{x,y} = a(x,y) - \overline{x}\overline{y} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)} y^{(i)} - \overline{x}\overline{y} = \frac{1}{n}\sum_{i=1}^{n} \left(x^{(i)} - \overline{x}\right)\left(y^{(i)}.\overline{y}\right)$ es la covarianza muestral y $s_x^2 = a(x,x) - (\overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n} \left(x^{(i)}\right)^2 - (\overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n} \left(x^{(i)} - \overline{x}\right)^2$ es la varianza muestral de x.

Supondremos que s_x^2 no es cero, es decir, que x presenta más de un valor. Si no, el sistema tendría infinitas soluciones.

Puede comprobarse que J es convexa y por lo tanto la solución que hemos obtenido de las ecuaciones normales empíricas es un mínimo local.

→ Además, como es único y J es continua, se trata del único mínimo global.

¿Es un mínimo local?

Las segundas derivadas parciales son

$$D_{1,1} = \frac{\partial^2}{\partial \theta_0^2} J(\theta_0, \theta_1) = \frac{1}{n}\sum_{i=1}^{n} 1 = 1$$
$$D_{1,2} = \frac{\partial^2}{\partial \theta_0 \theta_1} J(\theta_0, \theta_1) = \frac{1}{n}\sum_{i=1}^{n} x^{(i)} = \overline{x}$$
$$D_{2,2} = \frac{\partial}{\partial \theta_1^2} J(\theta_0, \theta_1) = \frac{1}{n}\sum_{i=1}^{n} \left(x^{(i)}\right)^2 = a(x,x)$$

Se verifica que $D_{1,1} = 1 > 0$ y si $D = (D_{i,j})$ es la matriz con esas derivadas, tenemos

$$|D| = \begin{vmatrix} 1 & \overline{x} \\ \overline{x} & a(x,x) \end{vmatrix} = a(x,x) - (\overline{x})^2 = s_x^2 > 0$$

por lo que el punto sería un mínimo local de J.

Como es el único mínimo local en $\mathbb{R}$ y J es continua, será el único mínimo global.
La solución óptima empírica coincide con la que obtendríamos sustituyendo en la solución teórica las medias, varianzas y covarianzas por sus estimaciones.

2.13) Regresión lineal múltiple

2.13.1) Modelo teórico

En el caso general queremos predecir Y (o X_{k+1}) a partir de k variables $X_1, \ldots, X_k$ (sobre el mismo espacio de probabilidad).

En el modelo lineal queremos construir una función

$$h_\theta(x_1, \ldots, x_k) \coloneqq \theta_0 + \theta_1 x_1 + \cdots + \theta_k x_K,$$

de forma que $h_\theta(X_1, \ldots, X_k)$ esté lo más cerca posible de Y.

Para medir el error usaremos de nuevo el error cuadrático medio

$$EMC(\theta) = E\left[(h_\theta(X_1, \ldots, X_k) - Y)^2\right].$$

Si consideramos una nueva variable constante (degenerada) $X_0 = 1$, podemos escribir ese error como

$$ECM(\theta) = E\left[(\theta' \mathbf{X} - Y)^2\right],$$

donde $\mathbf{X} = (X_0, X_1, \ldots, X_k)'$ y $\theta' = (\theta_0, \theta_1, \ldots, \theta_k) \in \mathbb{R}^{k+1}$.

2.13.2) Obtención del mínimo

Como en el caso $k = 1$ se puede comprobar que esta función es convexa por lo que tendrá un único minimo $\hat{\theta}' \in \mathbb{R}^{k+1}$.

Para detectarlo hacemos las derivadas parciales iguales a cero

$$\frac{\partial}{\partial \theta_j} ECM(\theta) = E[2(\theta' \mathbf{X} - Y) X_j] = 0$$

Obteniendo

$$\theta' E(\mathbf{X} X_j) = E(Y X_j),$$

para $j = 0, \ldots, k$, es decir,

$$\theta' E(\mathbf{X}\mathbf{X}') = E(Y \mathbf{X}')$$

O equivalente,

$$E(\mathbf{X}\mathbf{X}')\theta = E(\mathbf{X}Y).$$

Con lo que la solución es

$$\hat{\theta} = (E(\mathbf{X}\mathbf{X}'))^{-1} E(\mathbf{X}Y)$$

siempre que existe la inversa de la matriz simétrica

$$A = E(\mathbf{X}\mathbf{X}') = (E(X_i X_j))_{ij}$$

y donde (por convenio)

$$E(\mathbf{X}Y) = (E(X_0 Y), \ldots, E(X_k Y))'.$$

2.14) Coeficiente de correlación múltiple

- Definición

Si $\mathbf{Z} = (X_1, \ldots, X_k, Y)'$ es un vector aleatorio se llama coeficiente de correlación múltiple al cuadrado de Y respecto de

$\mathbf{X} = (X_1, \ldots, X_k)'$ a

$$\text{Corr}^2(\mathbf{X}, Y) = \rho^2_{k+1,(1,\ldots,k)} = \frac{v'_{1,2} V_{\mathbf{X}}^{-1} v_{1,2}}{\sigma^2_Y}$$

donde $\text{Cov}(\mathbf{Z}) = \begin{pmatrix} V_{\mathbf{X}} & v_{1,2} \\ v'_{1,2} & \sigma^2_Y \end{pmatrix}$, $V_{\mathbf{X}} = \text{Cov}(\mathbf{X})$, $\sigma^2_Y = \text{Var}(Y)$, y $v'_{1,2} = (\sigma_{1,k+1}, \ldots, \sigma_{k,k+1}) = \text{Cov}(X_1, Y), \ldots, \text{Cov}(X_k, Y)$.

Nota:
Si $k = 1$, es decir, si tenemos el vector aleatorio bidimensional (X, Y), entonces el **coeficiente de correlación múltiple al cuadrado** se corresponde con la **correlación de Pearson al cuadrado** (de X e Y), esto es,

$$\rho^2_{2,(1)} = \frac{\sigma_{1,2}\sigma^{-1}_{1,1}\sigma_{1,2}}{\sigma_{2,2}} = \rho^2_{1,2}$$

- Proposición

El coeficiente de correlación múltiple es el máximo de las correlaciones lineales al cuadrado de Y con combinaciones lineales de $\mathbf{X} = (X_1, \ldots, X_k)'$, es decir,

$$\max_{\alpha} \text{Corr}^2(Y, \alpha'\mathbf{X}) = \rho^2_{k+1,(1,\ldots,k)}$$

y ese máximo se obtiene con $\alpha = \lambda V_{\mathbf{X}}^{-1} v_{1,2}$, para $\lambda \neq 0$.

- Demostración

De la definición se tiene

$$\text{Corr}^2(Y, \alpha'\mathbf{X}) = \frac{(\text{Cov}(Y, \alpha'\mathbf{X}))^2}{\sigma^2_Y \text{Var}(\alpha'\mathbf{X})} = \frac{(\text{Cov}(Y, \mathbf{X})\alpha)^2}{\sigma^2_Y \text{Cov}(\alpha'\mathbf{X}, \alpha'\mathbf{X})}$$
$$= \frac{(\alpha' v_{1,2})^2}{\sigma^2_Y \alpha V_{\mathbf{X}} \alpha} = \frac{\left(\alpha' V_{\mathbf{X}}^{\frac{1}{2}} V_{\mathbf{X}}^{-\frac{1}{2}} v_{1,2}\right)^2}{\sigma^2_Y \alpha' V_{\mathbf{X}} \alpha}$$

Y usando la desigualdad de Cauchy-Schwarz, para $\mathbf{x}' = \alpha' V_{\mathbf{X}}^{\frac{1}{2}}$ e $y = V_{\mathbf{X}}^{-\frac{1}{2}} v_{1,2}$, se tiene

$$\text{Corr}^2(Y, \alpha'\mathbf{X}) \leq \frac{\alpha' V_{\mathbf{X}} \alpha v'_{1,2} V_{\mathbf{X}}^{-1} v_{1,2}}{\sigma^2_Y \alpha' V_{\mathbf{X}} \alpha} = \frac{v'_{1,2} V_{\mathbf{X}}^{-1} v_{1,2}}{\sigma^2_Y},$$

es decir, $\rho^2_{k+1,(1,\ldots,k)}$ es un cota superior.

Además, la igualdad en Cauchy-Schwarz se obtiene si y solo si los vectores $\mathbf{x}$ e $\mathbf{y}$ tienen la misma dirección

$$\mathbf{x} = V_{\mathbf{X}}^{-\frac{1}{2}} \alpha = \lambda \mathbf{y} = \lambda V_{\mathbf{X}}^{-\frac{1}{2}} v_{1,2},$$

es decir, si $\alpha = \lambda V_{\mathbf{X}}^{-1} v_{1,2}$ para $\lambda \neq 0$.

2.14.1) Desigualdad de Cauchy-Schwarz

Para vectores columna $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$, se verifica

$$(\mathbf{x}'\mathbf{y})^2 \leq (\mathbf{x}'\mathbf{x})(\mathbf{y}'\mathbf{y}),$$

y se obtiene la igualdad si y solo si los vectores $\mathbf{x}$ e $\mathbf{y}$ tienen la misma dirección, esto es, $\mathbf{x} = \lambda \mathbf{y}$.

2.14.2) Consecuencia

- Proposición

Si las variables $\mathbf{X} = (X_1, \ldots, X_k)'$ son independientes (o incorreladas) entre sí, entonces

$$\text{Corr}^2(\mathbf{X}, Y) = \sum_{j=1}^{k} \text{Corr}^2(X_j, Y).$$

- Demostración

La demostración es inmediata ya que si $\sigma_{i,j} = 0$ para $i \neq j$, $i, j \in \{1, \ldots, k\}$, $V_{\mathbf{X}}$ es diagonal, y se tiene

$$\text{Corr}^2(\mathbf{X}, Y) = \frac{v_{1,2}' V_{\mathbf{X}}^{-1} v_{1,2}}{\sigma_Y^2} = \sum_{j=1}^{k} \frac{\sigma_{j,k+1}}{\sigma_Y^2 \sigma_{j,j}} = \sum_{j=1}^{k} \rho_{j,k+1}^2.$$

Note que si sustituimos X_i e Y por $X_i - \mu_i$ e $Y - \mu_Y$ en la expresión de $\hat{\theta}$, podemos eliminar X_0 (como la solución pasa por el vectore de medias, en este caso $\theta_0 = 0$), la correlación no varía y tenemos

$$\hat{\theta} = \{E[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{X} - \mu_{\mathbf{X}})']\}^{-1} E[(\mathbf{X} - \mu_{\mathbf{X}})(Y - \mu_Y)] = V_{\mathbf{X}}^{-1} \sigma_{1,2},$$

es decir, ambas soluciones coinciden.

Así, podemos predecir Y usando

$$Y - \mu_Y = \text{Cov}(Y, \mathbf{X}) V_{\mathbf{X}}^{-1} (\mathbf{X} - \mu_{\mathbf{X}}),$$

es decir,

$$h_{\hat{\theta}}(x) = \mu_Y + \text{Cov}(Y, \mathbf{X}) V_{\mathbf{X}}^{-1} (\mathbf{x} - \mu_{\mathbf{X}})$$

donde $\text{Cov}(Y, \mathbf{X}) = (\text{Cov}(Y, X_1), \ldots, \text{Cov}(Y, X_k))$.

2.15) Selección de variables

En algunos casos podemos querer detectar las variables del conjunto $X_1, \ldots, X_k$ que mejor predicen Y.

2.15.1) Una opción

Seleccionar la variable $Z_1 = X_{j1}$ que maximice la correlación al cuadrado con Y:

$$j_1 = \max_{j=1,\ldots,k} \text{Corr}^2(X_j, Y).$$

Calcular la recta de regresión h_1 basada en Z_1 y el residuo $R_1 = h_1(Z_1) - Y$.

Seleccionar la variable $Z_2 = X_{j2}$ con $j \neq j_1$ que más información tenga sobre ese residuo:

$$j_2 = \max_{j=1,\ldots,k, j \neq j_1} \text{Corr}^2(X_j, R_1).$$

Calcular el segundo residuo $R_2 = h_2(Z_2) - R_1$.

Continuar así hasta obtener el número de variables deseado o hasta que la correlacón múltiple sea tan grande como se desee.

2.15.2) Otra opción

Fijar de antemano el número de variables deseadas $p < k$.

Calcular las correlaciones múltiples de todos los subconjuntos con p variables.

Seleccionar al que tenga una mayor correlación múltiple.

2.15.3) Otra opción más sencilla

Considerar desde el inicio vairables estandarizadas.

Calcular $\hat{\theta}^*/$ para estas variables.

Seleccionar primero las que tengan un mayor coeficiente $\hat{\theta}_j^*$ en valor absoluto.

- Son las que más influyen en el valor Y ya que todas las variables tienen magnitudes similares.

2.16) Inferencia y predicción

¡No se conocen los valores teóricos!

En la práctica los valores teóricos deben ser estimados:

- Una primera opción:
  - Seleccionar una muestra aleatoria simple.
  - Estimar las medias, varianzas y cuasivarianzas y usarlas para estimar sus respectivos valores teóricos en las expresiones obtenidas en la sección anterior.
- Otra opción:
  - Considerar el problema empírico.
  - Partiremos de una muestra (training sample): $\left(x_1^{(i)}, \ldots, x_k^{(i)}, y^{(i)}\right)$, para $i = 1, \ldots, n$, donde concoemos los valores de y para esos valores de x.
  - Los datos se colocarán como $k+1$ columnas (variables) y n filas (objetos o individuos).
  - Cada variable (sus datos) también se puede ver como un punto de $\mathbb{R}^n$.

2.17) Función costo empírica

Función de predicción lineal será:

$$h_\theta(x) \coloneqq \theta' x = \theta_0 + \theta_1 x_1 + \cdots + \theta_k x_k,$$

donde $\theta = (\theta_0, \ldots, \theta_k)'$ y $x = (x_0, \ldots, x_k)'$.

Para simplificar la notación es conveniente añadir una variable (columna) x_0 con n unos.

La matriz de datos para x se representará como $M = (m_{i,j}) = \left(x_j^{(i)}\right)$, para $i = 1, \ldots, n$ (fila) y $j = 0, \ldots, k$ (columna).
Objetivo: Minimizar la función coste (proporcional al error cuadrático medio)

$$J(\theta) \coloneqq \frac{1}{2n} \sum_{i=1}^{n} \left(h_\theta(\mathrm{x}^{(i)}) - y^{(i)}\right)^2 = \frac{1}{2n} \sum_{i=1}^{n} \left(\theta' \mathbf{x}^{(i)} - y^{(i)}\right)^2,$$

donde $\mathbf{x}^{(i)} = \left(x_0^{(i)}, \ldots, x_k^{(i)}\right)$ e $y^{(i)}$ representan las medidas del individuo i-ésimo.

Función en forma matricial:

$$J(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \left(\theta' \mathbf{x}^{(i)} - y^{(i)}\right)^2 = \frac{1}{2n} (M\theta - y)'(M\theta - y),$$

siendo $y = \left(y^{(1)}, \ldots, y^{(n)}\right)$.

Alternativamente,

$$J(\theta) = \frac{1}{2n}(M\theta - y)'(M\theta - y) = \frac{1}{2n}(\theta' M' M\theta - 2\theta' M' y + y' y).$$

De nuevo tenemos una función J convexa y para detectar su valor mínimo haremos las derivadas parciales y trataremos de resolver

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(\theta' x^{(i)} - y^{(i)} \right) x_j^{(i)} = 0,$$

para $j = 0, 1, \ldots, k$.

Lo que es equivalente a

$$\frac{1}{n}(\theta' M' - y')M = 0.$$

Equivalente también a las denominadas ecuaciones normales

$$M' M\theta - M' y = 0,$$

siendo 0 el vector de ceros con la dimensión adecuada.

Por lo tanto la solución es

$$\hat{\theta} = (M' M)^{-1} M' y$$

siempre que exista la inversa de $M'M$.

Existe la inversa de $M'M$

Esta inversa puede no existir:

- Porque haya pocos datos ($n < k$).
- Porque algunas variables sean dependientes (por ejemplo, si $X_2 = \lambda X_1$).

En esos casos la solución no es única.

Para evitarlos debemos:

- Tomar más datos en el primer caso.
- Eliminar variables redundantes en el segundo.

Como la matriz de datos M es una matriz $n \times k$, $M'M$ es una matriz $k \times k$.

- Si el número de variables, k, es muy grande (mayor que 10000).
  - Podemos tener problemas al calcular su inversa.
  - En esos casos usaremos el algoritmo gradiente descendiente para J.

2.17.1) Descomposición de la variabilidad

- Variabilidad total: $SCT = \sum_{i=1}^{n} (y_i - \overline{y})^2$ (suma de cuadrados total).
- Podemos descomponer la variabilidad total en dos sumandos:

$$SCT = SCE + SCR$$

- SCE es la variabilidad explicada por la regresión: $SCE = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$ (suma de cuadrados explicada).

- SCR es la variabilidad no explicada por la regresión: $SCR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ (suma de cuadrados residual).

2.17.2) Coeficiente de determinación: R^2

El coeficiente de determinación se define como la proposición de variabilidad de la variable dependiente que es explicada por los regresores:

$$R^2 = \frac{SCE}{SCT} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}; \qquad R^2 = 1 - \frac{SCR}{SCT} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

Análogamente,

$$1 - R^2 = \frac{SCR}{SCT} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

indicaría la parte de Y no explicada por los regresores y que se queda en el residuo.

2.17.3) Propiedades

- $0 \leq R^2 \leq 1$.
- Cuando $R^2 = 1$ existe una relación exacta entre los valores ajustados y la variable respuesta.
- Cuando $R^2 = 0$, $\hat{y}_i = \overline{y}$, para todo $i = 1, \ldots, n$.
- R^2 coincide con el coeficiente de correlación múltiple al cuadrado entre y y las k variables regresoras.
  - En regresión lineal simple $R^2 = \rho^2$, donde ρ es el coeficiente de correlación lineal.
- Además $R^2 = \rho^2_{y,\hat{y}'}$, es decir, R^2 coincide con el coeficiente de correlación lineal simple entre las variables y e $\hat{y}$.

El coeficiente de determinación presenta el inconveniente de aumentar siempre que aumenta el número de variables regresoras (algunas veces de forma artificial).

2.17.4) Coeficiente de determinación ajustado

Para penalizar el número de variables regresoras que se incluyen en el modelo de regresión, es conveniente utilizar el coeficiente de determinación corregido por el número de grados de libertad, denominado coeficiente de determinación ajustado, definido como:

$$\overline{R}^2 = 1 - \frac{\frac{SCR}{n-k-1}}{\frac{SCT}{n-1}}$$

O equivalente,

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

2.18) Extensiones del modelo de regresión múltiple

2.18.1) Planteamiento

El modelo de regresión lineal se puede usar para añadir variables a nuestro modelo inicial (univariante o multivariante).

El modelo más típico es el modelo polinómico:

$$h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_g x^g,$$

donde el entero g representa el grado del polinomio que consideremos más adecuado.

- En el modelo de regresión lineal multivariante consideraremos:

$$X_1 = X, X_2 = X^2, \ldots, X_g = X^g$$

Otro ejemplo: Si queremos predecir Y a partir de X_1, X_2 y X_3, pero el modelo lineal no funciona bien, podemos considerar el modelo:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_1 x_2 + \theta_5 x_1 x_3 + \theta_6 x_2 x_3.$$

Las gráficas bidimensionales (nubes de puntos) nos pueden dar una idea de las relaciones que pueden mejorar nuestras estimaciones.

2.18.2) Problema de sobreajuste (`overfitting`)

Es evidente que aumentando el grado de la regresión polinómica, disminuirá el valor de J.

Si no hay más de un dato para cada x y consideramos $g = n$ podemos conseguir un ajuste prefecto (interpolación polinómica).

- Sin embargo, este ajuste perfecto para los datos de la muestra de entrenamiento, no tiene por qué funcionar mejor cuando lo usemos en otros datos.
- De hecho, casi siempre funciona peor.

También se nos puede dar el caso contrario, denominado subajuste (`underfitting`)

Un caso sencillo

Para ilustrar este hecho consideremos un caso sencillo en el que ajustamos a los datos una recta (`underfitting`), un polinomio cúbico y un polinomio de orden $g = n$ (`overfitting`).

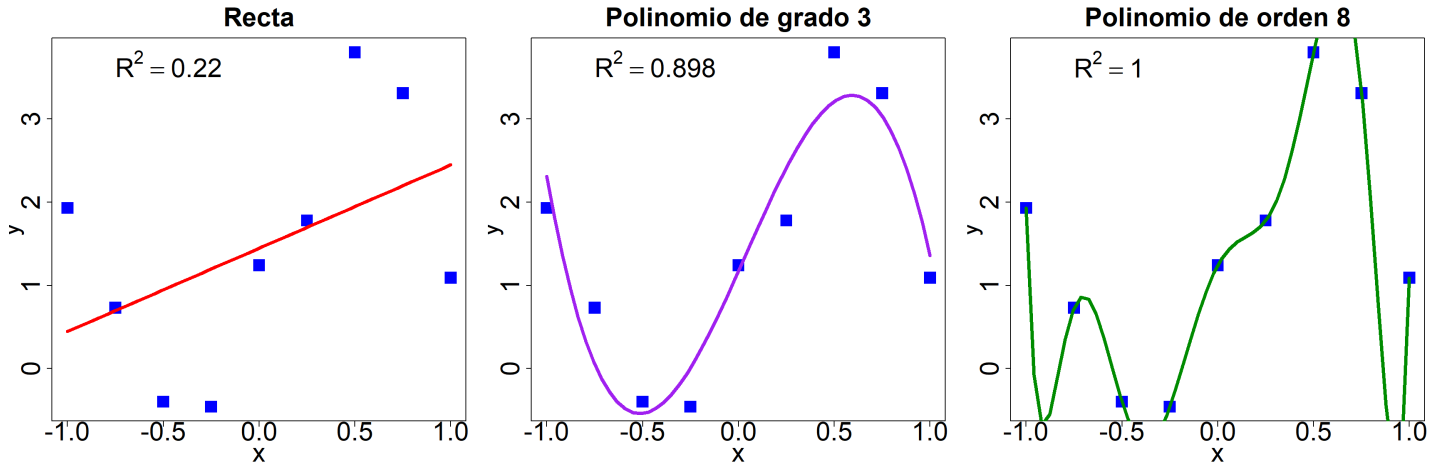
```
par(mfrow = c(1,3))
x = seq(-1, 1, 0.25)
y = c(1.93, 0.73, -0.40, -0.46, 1.24, 1.78, 3.80, 3.31, 1.09)

model1 = lm(y ~ x)
r2.model1 = summary(model1)$r.squared
model2 = lm(y ~ x + I(x^2) + I(x^3))
r2.model2 = summary(model2)$r.squared
model3 = lm(y ~ x + I(x^2) + I(x^3) +  I(x^4) + I(x^5) +  I(x^6) + I(x^7) + I(x^8))
r2.model3 = summary(model3)$r.squared

x1 = seq(-1, 1, length = 50)

plot(x, y, type = "p", col = "blue", cex = 3, pch = 15, main = "Recta", cex.lab = 3,
    cex.axis = 3, cex.main = 3)
lines(x1, predict(model1, newdata = data.frame(x = x1)), col = "red", lwd = 4, lty =
    1, cex = 0.2)
```

```
text(-0.75, 3.5, bquote(R^2 == .(format(r2.model1, digits = 3))), adj = c(0, 0), cex
    = 3)
plot(x, y, type = "p", col = "blue", cex = 3, pch = 15, main = "Polinomio de grado 3"
    , cex.lab = 3, cex.axis = 3, cex.main = 3)
lines(x1, predict(model2, newdata = data.frame(x = x1)), col="purple", lwd=4, lty =
    1, cex = 0.2)
text(-0.75, 3.5, bquote(R^2 == .(format(r2.model2, digits = 3))), adj = c(0, 0), cex
    = 3)
plot(x, y, type = "p", col ="blue", cex = 3, pch = 15, main = "Polinomio de orden 8",
     cex.lab = 3, cex.axis = 3, cex.main = 3)
lines(x1, predict(model3, newdata = data.frame(x = x1)), col = "green4", lwd = 4, lty
     = 1, cex = 0.2)
text(-0.75, 3.5, bquote(R^2 == .(format(r2.model3, digits = 3))), adj = c(0, 0), cex
    = 3)
```



Otro ejemplo

Ajustamos a unos datos una recta, una parábola y un polinomio cúbico.

¿Realmente se mejora el ajuste?

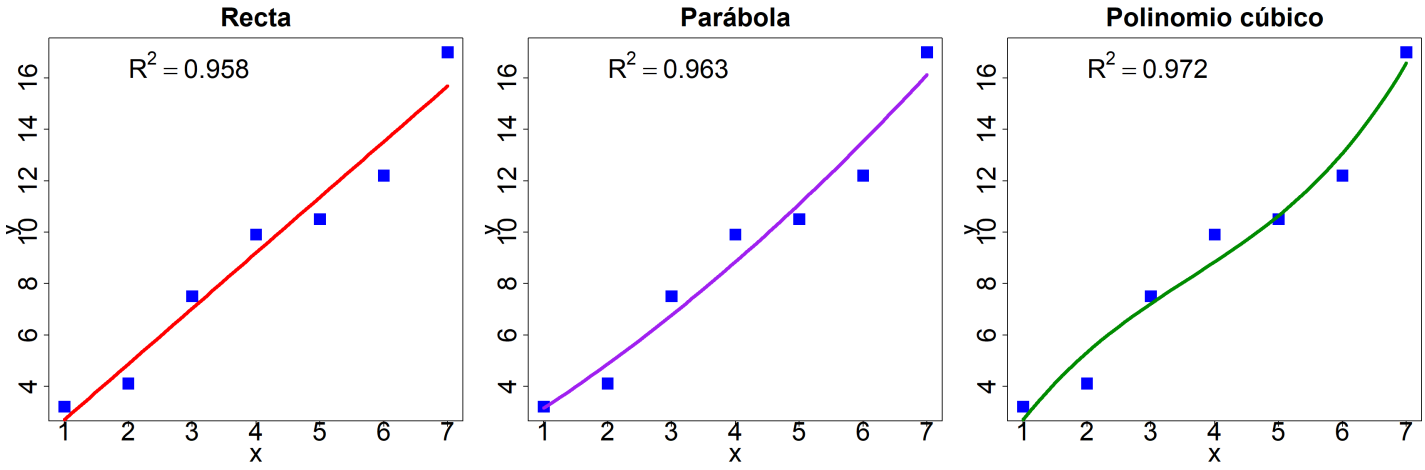
```
par(mfrow = c(1,3))
x = 1:7
y = c(3.2, 4.1, 7.5, 9.9, 10.5, 12.2, 17)

model1 = lm(y ~ x)
r2.model1 = summary(model1)$r.squared
model2 = lm(y ~ x + I(x^2))
r2.model2 = summary(model2)$r.squared
model3 = lm(y ~ x + I(x^2) + I(x^3))
r2.model3 = summary(model3)$r.squared

x1 = seq(1, 7, length = 50)

plot(x, y, type = "p", col = "blue", cex = 3, pch = 15, main = "Recta", cex.lab = 3,
    cex.axis = 3, cex.main = 3)
```

```
lines(x1, predict(model1, newdata = data.frame(x = x1)), col = "red", lwd = 4, lty = 1, cex = 0.2)
text(2, 16, bquote(R^2 == .(format(r2.model1, digits = 3))), adj = c(0, 0), cex = 3)
plot(x, y, type = "p", col = "blue", cex = 3, pch = 15, main = "Parábola", cex.lab = 3, cex.axis = 3, cex.main = 3)
lines(x1, predict(model2, newdata = data.frame(x = x1)), col="purple", lwd=4, lty = 1, cex = 0.2)
text(2, 16, bquote(R^2 == .(format(r2.model2, digits = 3))), adj = c(0, 0), cex = 3)
plot(x, y, type = "p", col ="blue", cex = 3, pch=15, main = "Polinomio cúbico", cex.lab = 3, cex.axis = 3, cex.main = 3)
lines(x1, predict(model3, newdata = data.frame(x = x1)), col = "green4", lwd = 4, lty = 1, cex = 0.2)
text(2, 16, bquote(R^2 == .(format(r2.model3, digits = 3))), adj = c(0, 0), cex = 3)
```



Para nuestro ejemplo

¿Cómo se comporta el coeficiente de determinación ajustado?

¿Qué modelo seleccionaría?

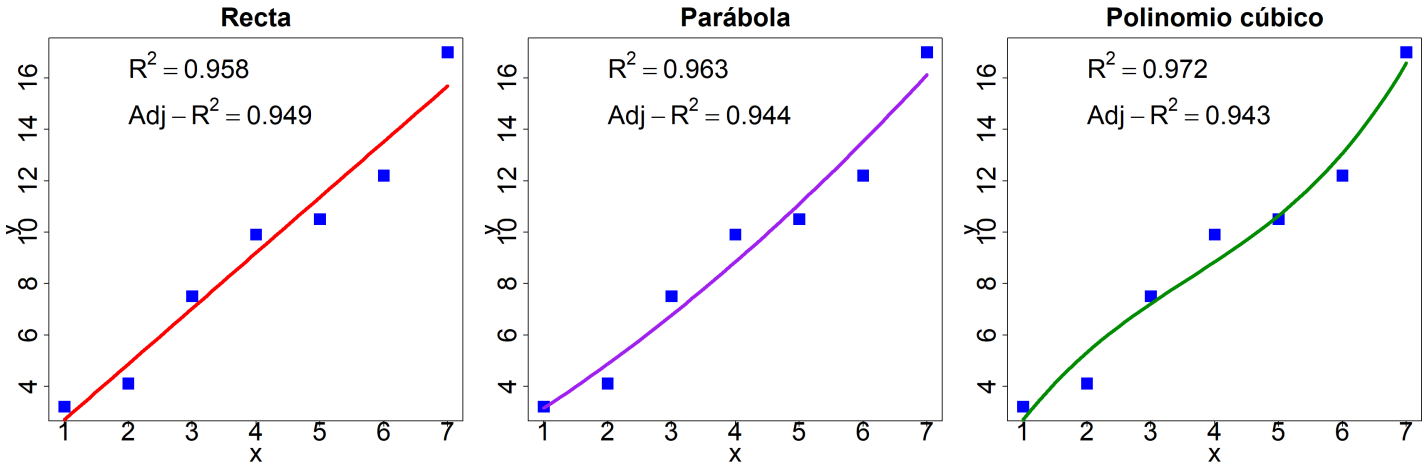
```
par(mfrow = c(1,3))
x = 1:7
y = c(3.2, 4.1, 7.5, 9.9, 10.5, 12.2, 17)

model1 = lm(y ~ x)
r2.model1 = summary(model1)$r.squared
adj.r2.model1 = summary(model1)$adj.r.squared
model2 = lm(y ~ x + I(x^2))
r2.model2 = summary(model2)$r.squared
adj.r2.model2 = summary(model2)$adj.r.squared
model3 = lm(y ~ x + I(x^2) + I(x^3))
r2.model3 = summary(model3)$r.squared
adj.r2.model3 = summary(model3)$adj.r.squared

x1 = seq(1, 7, length = 50)

```

```r
plot(x, y, type = "p", col = "blue", cex = 3, pch = 15, main = "Recta", cex.lab = 3, cex.axis = 3, cex.main = 3)
lines(x1, predict(model1, newdata = data.frame(x = x1)), col = "red", lwd = 4, lty = 1, cex = 0.2)
text(2, 16, bquote(R^2 == .(format(r2.model1, digits = 3))), adj = c(0, 0), cex = 3)
text(2, 14, bquote(Adj-R^2 == .(format(adj.r2.model1, digits = 3))), adj = c(0, 0), cex = 3)
plot(x, y, type = "p", col = "blue", cex = 3, pch = 15, main = "Parábola", cex.lab = 3, cex.axis = 3, cex.main = 3)
lines(x1, predict(model2, newdata = data.frame(x = x1)), col="purple", lwd=4, lty = 1, cex = 0.2)
text(2, 16, bquote(R^2 == .(format(r2.model2, digits = 3))), adj = c(0, 0), cex = 3)
text(2, 14, bquote(Adj-R^2 == .(format(adj.r2.model2, digits = 3))), adj = c(0, 0), cex = 3)
plot(x, y, type = "p", col ="blue", cex = 3, pch = 15, main = "Polinomio cúbico", cex.lab = 3, cex.axis = 3, cex.main = 3)
lines(x1, predict(model3, newdata = data.frame(x = x1)), col = "green4", lwd = 4, lty = 1, cex = 0.2)
text(2, 16, bquote(R^2 == .(format(r2.model3, digits = 3))), adj = c(0, 0), cex = 3)
text(2, 14, bquote(Adj-R^2 == .(format(adj.r2.model3, digits = 3))), adj = c(0, 0), cex = 3)
```



¿Cómo evitar estos problemas?

Procedimiento para muestras grandes

¿Cómo elegir el número óptimo de variables? ¿Y cúales son las más adecuadas?

- Separaremos nuestra muestra (de forma aleatoria) en dos grupos.
  - El primer grupo se usará para estimar los coeficientes óptimos para cada g.
  - El segundo grupo se utilizará para calcular los errores cuadráticos medios para cada g.
    - → Obviamente, escogeremos el grado (o grupo de variables) con menor error.
    - → De nuevo ese error nos dará una estimación menor del error real que se obtendrá con ese g óptimo.
- Para hacernos una idea del error real deberemos guardar un tercer grupo de datos para calcular el error en ellos.

El número de datos en cada grupo dependen de muchos factores:

- Tamaño muestral n.
- Número de variables consideradas k.
- Tiempo de programación, etc.

Por ejemplo, si tenemos $n = 100$ datos, podríamos dividir el conjunto en:

- Un subconjunto con 60 datos para el cálculo de h.
- Un subconjunto con 20 datos para determinar el g óptimo.
- Un subconjunto con los otros 20 para estimar el error real en las predicciones futuras.

Si nuestra muestra tienen pocos datos, para aplicar este procedimiento deberemos aplicarlo a cada dato eliminándolo del procedimiento para estimar h.

RELACIÓN DE PROBLEMAS: REGRESIÓN LINEAL SIMPLE Y MÚLTIIPLE
ANÁLISIS ESTADÍSTICO MULTIVARIANTE
GRADO EN CIENCIA E INGENIERÍA DE DATOS

1. Dado el vector aleatorio (X, Y) con función de densidad

$$f(x,y) = \begin{cases} 2 & \text{si } 0 < x < 1, 0 < y < x \\ 0 & \text{en otro caso} \end{cases}$$

   a) Obtener la curva de regresión para predecir Y en función de valores de la variable X.

   b) ¿Coincide con la recta de regresión?

2. Sea (X, Y) un vector aleatorio con función de densidad

$$f(x,y) = \begin{cases} \frac{3}{4}\left[xy + \frac{x^2}{2}\right] & \text{si } 0 < x < 1, 0 < y < 2 \\ 0 & \text{en otro caso} \end{cases}$$

   A partir de la distribución de Y condicionada a $X = x$, obtener la curva de regresión para predecir Y a partir de valores de X y proporcionar una predicción para $X = 2/3$.

3. Sabiendo que el vector (X, Y) tiene una distribución normal con medias 1 y 2, varianzas 2 y correlación -1/2, calcular la curva de regresión para predecir Y a partir de valores de X y obtener una predicción para $X = 1.5$.

4. Encontrar la recta de regresión para el conjunto de datos:

$$\{(x_i, y_i)\} = \{(1,4),\ (2,2),\ (1,5),\ (5,3),\ (6,2)\}$$

   Estimar y para $x = 3$. ¿Será fiable esa aproximación?

5. El rendimiento de una reacción química depende de la concentración del reactivo y de la temperatura de la operación.

Rendimiento	Concentración	Temperatura
81	1.00	150
89	1.00	180
83	2.00	150
91	2.00	180
79	1.00	150
87	1.00	180
84	2.00	150
90	2.00	180

   En el modelo de regresión del rendimiento sobre la temperatura y la concentración, los valores ajustados son

$$\hat{\mathbf{y}} = (\ 80.25,\ \ 87.75,\ \ 83.25,\ \ 90.75,\ \ 80.25,\ \ 87.75,\ \ 83.25,\ \ 90.75\ )'$$

   Calcular el error cuadrático medio y el coeficiente de determinación e interpretar su valor.

6. Un modelo ajustado para predecir la extracción de manganeso en % (y) a partir del tamaño de partícula en mm (x_1), la cantidad de dióxido de azufre en múltiplos de la cantidad estequiométrica necesaria para la disolución de manganeso (x_2) y la duración de la filtración en minutos (x_3) están dadas como

$$y = 56.145 - 9.0469x_1 - 33.421x_2 + 0.243x_3 - 0.5963x_1x_2$$
$$-0.0394x_1x_3 + 0.60022x_2x_3 + 0.6901x_1^2 + 11.7244x_2^2 - 0.0097x_3^2$$

Se consideraron 27 observaciones, con $\sum_i^n (y_i - \hat{y}_i)^2 = 209.55$, $\sum_i^n (y_i - \overline{y})^2 = 6777.5$.

Se pide:

a) Obtener una predicción para el porcentaje de extracción cuando el tamaño de partícula es 3 mm, la cantidad de dióxido de azufre 1.5 y la duración de la filtración es de 20 minutos.

b) ¿Es posible predecir el cambio en el porcentaje de extracción cuando la duración de la filtración aumenta en un minuto? Si la respuesta es afirmativa, encontrar el cambio pronosticado. Si la respuesta es negativa, ¿qué otra información se necesitaría para determinarlo?

c) Calcular el coeficiente de determinación R^2. Interpretar su valor.

7. Se efectúa un estudio sobre el desgaste de un cojinete (y) y su relación con la viscosidad del aceite (x_1) y carga (x_2). Se obtienen los datos siguientes:

y	x_1	x_2
193	1.6	851
230	15.5	816
172	22.0	1058
91	43.0	1201
113	33.0	1357
125	40.0	1115

a) El análisis de regresión lineal múltiple con los datos de la tabla anterior proporciona $\hat{\theta}$=(350.994, -1.27999,-0.15390)' y los valores ajustados son

$$\hat{\mathbf{y}}\text{=(217.99, 205.69, 160.18, 111.46, 100.17, 128.51)'}$$

Calcular una estimación de la varianza residual y el coeficiente de determinación R^2.

b) Utilizar el modelo para predecir el desgaste cuando $x_1 = 25$ y $x_2 = 1000$.

8. En un proceso industrial se sospecha que la dureza de las láminas de acero reducido en frío depende del contenido en cobre (x_1) y de la temperatura de recocido (x_2). Para comprobar esta suposición se mide en doce especímenes de láminas de acero la dureza para varios valores del contenido de cobre y de la temperatura, obteniéndose los siguientes resultados:

Dureza (Rockwell 30-T)	Contenido de cobre (%) (x_1)	Temperatura de recocido (? F) (x_2)
79.1	0.025	1000
65.3	0.025	1100
55.5	0.025	1200
56.6	0.025	1300
81.1	0.15	1000
69.9	0.15	1100
57.6	0.15	1200
55.6	0.15	1300
85.5	0.2	1000
72.0	0.2	1100
60.9	0.2	1200
59.1	0.2	1300

a) Plantear el modelo de regresión lineal múltiple para explicar la dureza de la lámina en función del contenido de cobre y de la temperatura.

b) Sabiendo que $\hat{\theta} = (161.404, 27.1923, -0.08547)'$, obtener una estimación para la varianza del error y el valor de R^2. Comentar la bondad del ajuste.

c) Se decide eliminar la variable x_1 del modelo y ajustar un modelo de regresión lineal simple para predecir la dureza en función de x_2.

1) ¿Qué modelo se obtendría?
2) ¿Qué cambio se espera en la dureza promedio si la temperatura se incrementa en 100 grados Farenheith?

1) Dado el vector aleatorio (X, Y) con función de densidad

$$f(x,y) = \begin{cases} 2 & \text{si } 0 < x < 1,\ 0 < y < x \\ 0 & \text{en otro caso} \end{cases}$$

a) Obtener la curva de regresión para predecir Y en función de valores de la variable X.

Primero saco las distribución marginal f_X:

$$\underset{0<x<1}{f_X(x)} = \int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}y = \int_0^1 2\,\mathrm{d}y = [2y]_0^1 = 2 \longrightarrow \begin{cases} 2 & \text{si } 0 < y < 1 \\ 0 & \text{en caso contrario} \end{cases}$$

$$\underset{0<x<1}{f_{Y|X}(y|x)} = \frac{f(x,y)}{f(x)} = \begin{cases} \dfrac{1}{x} & \text{si } 0 < y < x \\ 0 & \text{en caso contrario} \end{cases}$$

Curva de regresión: $h_{\text{opt}}(X) = E[Y|X=x] = \displaystyle\int_{-\infty}^{+\infty} y \cdot f_{Y|X}(y|x)\,\mathrm{d}y = \int_0^x y \cdot \frac{1}{x}\,\mathrm{d}y = \frac{1}{x}\left[\frac{y^2}{2}\right]_{y=0}^{y=x} = \frac{1}{x}\cdot\frac{x^2}{2} = \frac{x}{2}$

b) ¿Coincide con la recta de regresión?

Recta de regresión de $Y|X=x$: $Y - \mu_Y = \dfrac{\cos(X,Y)}{\sigma_X^2}(x - \mu_X) \longrightarrow y - \dfrac{1}{3} - \dfrac{\frac{1}{36}}{\frac{1}{18}}\left(x - \dfrac{2}{3}\right) \longrightarrow \boxed{y = \dfrac{x}{2}}$

$$E[X] = \int_0^1 2x\,\mathrm{d}x = \left[2\cdot\frac{x^3}{3}\right]_{x=0}^{x=1} = \frac{2}{3}$$

$$E[Y] = \int_0^1 y\cdot 2(1-y)\,\mathrm{d}y = 2\int_0^1 (y - y^2)\,\mathrm{d}y = 2\left[\frac{y^2}{2} - \frac{y^3}{3}\right]_{y=0}^{y=1} = 2\cdot\frac{1}{6} = \frac{1}{3}$$

$$\sigma_X^2 = E[X^2] - (E[X])^2 = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{2} - \frac{4}{9} = \frac{1}{18}$$

$$E[X^2] = \int_0^1 x^2\cdot 2x\,\mathrm{d}x = 2\int_0^1 x^3\,\mathrm{d}x = 2\left[\frac{x^4}{4}\right]_{x=0}^{x=1} = \frac{1}{2}$$

$$\mathrm{Cov}(X,Y) = E[X\cdot Y] - E[X]\cdot E[Y] = \frac{1}{4} - \frac{2}{3}\cdot\frac{1}{3} = \frac{1}{36}$$

$$E[X\cdot Y] = \int_0^1\int_0^x 2xy\,\mathrm{d}y\,\mathrm{d}x = 2\int_0^1 x\left[\frac{y^2}{2}\right]_{y=0}^{y=x}\mathrm{d}x = 2\int_0^1 x\cdot\frac{x^2}{2}\,\mathrm{d}x = 2\cdot\left[\frac{x^4}{4}\right]_{x=0}^{x=1} = \frac{1}{4}$$

c) $\rho_{X,Y}^2$, $\mathrm{Var}(R)$, ECM

2) Sea (X, Y) un vector aleatorio con función de densidad

$$f(x,y) = \begin{cases} \dfrac{3}{4}\left[xy + \dfrac{x^2}{2}\right] & \text{si } 0 < x < 1,\ 0 < y < 2 \\ 0 & \text{en otro caso} \end{cases}$$

A partir de la distribución de Y condicionada a $X = x$, obtener la curva de regresión para predecir Y a partir de los valores de X y proporcionar una predicción para $X = \dfrac{2}{3}$

$$h_{\text{opt}}(x) = E[Y|X=x] = \int_{-\infty}^{+\infty} y\cdot f_{Y|X}(y|x)\,\mathrm{d}y = \int_0^2 y\cdot\frac{y + \frac{x}{2}}{x+2}\,\mathrm{d}y = \frac{1}{x+2}\int_0^2 y\cdot(y + \frac{x}{2})\,\mathrm{d}y = \frac{1}{x+2}\cdot\left[\frac{y^3}{3} - \frac{y^2x}{4}\right]_0^2 = \frac{1}{x+2}\left(\frac{8}{3} - x\right)$$

$$f_X(x) = \begin{cases} \dfrac{3}{4}(x^2 + 2x) & \text{si } 0 < x < 1 \\ 0 & \text{en otro caso} \end{cases}$$

$$f_{\substack{Y|X\\0<x<1}}(y|x) = \begin{cases} \dfrac{y+\frac{x}{2}}{x+2} & \text{si } 0<y<2 \\ 0 & \text{en otro caso} \end{cases}$$

Para $x = \dfrac{2}{3}$, $h_{\text{opt}}\left(\dfrac{2}{3}\right) = \dfrac{1}{\frac{2}{3}+2} \cdot \left(\dfrac{8}{3}+\dfrac{2}{3}\right) = \boxed{1.25}$

3) Sabiendo que el vector (X,Y) tiene una distribución normal con medias 1 y 2, varianzas 2 y correlación $-\dfrac{1}{2}$, calcular la curva de regresión para predecir Y a partir de valores de X y obtener una predicción para $X = 1.5$

$(X,Y) \rightsquigarrow \mathcal{N}_2(\mu, V)$

$\mu = (1,2)$

$V = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ $\qquad f_{X,Y} = -\dfrac{1}{2} = \dfrac{\text{Cov}(X,Y)}{\sqrt{\sigma_X^2 \cdot \sigma_Y^2}} \longrightarrow \text{Cov}(X,Y) = -\dfrac{1}{2} \cdot \sqrt{2 \cdot 2} = -1$

$$h_{\text{top}} = \mu_Y + \frac{\text{Cov}(X,Y)}{\sigma_X^2}(x-\mu_X) = 2 + \frac{-1}{2}(x-1) = -\frac{x}{2} + \frac{5}{2} \longrightarrow h_{\text{opt}}(1.5) = 2 - \frac{1}{2} \cdot \frac{1}{2} = \boxed{\frac{7}{4}}$$

4) Encontrar la recta de regresión para el conjunto de datos:

$$\{(x_i, y_i)\} = \{(1,4),(2,2),(1,5),(5,3),(6,2)\}$$

Estimar y para $x = 3$. ¿Será fiable esa aproximación?

Recta de regresión: $y = \hat{\theta}_0 + \hat{\theta}_1 x$

$$\hat{\theta}_0 = \overline{y} - \hat{\theta}_1 \overline{x} = 3.2 - (-0.\overline{36}) \cdot 3 = 4.29$$

$$\hat{\theta}_1 = \frac{S_{xy}}{S_x^2} = -\frac{1.6}{4.4} = -0.\overline{36}$$

$n = 5$

$\overline{x} = 3$

$\overline{y} = 3.2$

$S_x^2 = \overline{x^2} - \overline{x}^2 = \dfrac{67}{3} - 3^2 = 13.4 - 9 = 4.4$

$S_{xy} = \overline{xy} - \overline{x} \cdot \overline{y} = 8 - 3 * 3.2 = -1.6$

Recta de regresión: $y = 4.29 - 0.\overline{36}x \xrightarrow{x=3} y = 4.29 - 0.\overline{36} \cdot 3 = 3.2$

Coeficiente de correlación lineal al cuadrado:

$$r_{xy}^2 = \frac{S_{xy}^2}{s_x^2 \cdot s_y^2} = \frac{(-1.6)^2}{4.4 \cdot 1.36} = 0.428 \text{ Ajuste malo}$$

$s_y^2 = \overline{y^2} - \overline{y}^2 = \dfrac{58}{5} - 3.2^2 = 1.36$

5) En el rendimiento de una reacción química depende de la concetración del reactivo y de la temperatura de la operación.

Rendimiento	Concentración	Temperatura
81	1.00	150
89	1.00	180
83	2.00	150
91	2.00	180
79	1.00	150
87	1.00	180
84	2.00	150
90	2.00	180

En el modelo de regresión del rendimiento sobre la temperatura y la concentración, los valores ajustados son

$$\hat{y} = \begin{pmatrix} 80.25, & 87.75, & 83.25, & 90.75, & 87.75, & 83.25, & 90.75 \end{pmatrix}'$$

Calcular el error cuadrático medio y el coeficiente de determinación e interpretar su valor.

Rend.(y)	$\hat{y}$	$y - \hat{y}$
81	80.25	0.75
89	87.75	1.25
83	83.25	-0.25
91	90.75	0.25
79	80.25	-1.25
87	87.75	-0.75
84	83.25	0.75
90	90.75	-0.75

$$ECM = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{5.5}{8} = \boxed{0.6875}$$

$$R^2 = 1 - \frac{SCR}{SCT} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - \frac{5.5}{136} = \boxed{0.9595}$$ Ajuste muy bueno

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 85.5$$

$$R^2 = \frac{SCE}{SCT} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

6) Un modelo ajustado para predecir la extracción de manganeso en % (y) a partir del tamaño de partícula en mm (x_1), la cantidad de dióxido de azufre en múltiplos

a) $\hat{y}_{x_1=3,x_2=1.5,x_3=20} = 56.145 - 9.0469 \cdot 3 - 33.421 \cdot 1.5 + 0.243 \cdot 20 - 0.5963 \cdot 3 \cdot 1.5 - 0.0394 \cdot 3 \cdot 20 + 0.60022 \cdot 1.5 \cdot 20 + 0.6901 \cdot 3^2 + 11.7244 \cdot (1.5)^2 - 0.0097 \cdot 20^2 = 25.4028$

b) $x_3 \longrightarrow x_3 + 1$

$$\begin{aligned}\hat{y}_{x_1,x_2,x_3+1} - \hat{y}_{x_1,x_2,x_3} &= 56.145 - 9.0496 \cdot x_1 - \cdots - 0.0097 \cdot (x_3+1)^2 - 56.145 - 9.0496 \cdot x_1 - \cdots - 0.0097 \cdot (x_3)^2 \\ &= 0.243 \cdot (x_3+1) - 0.0394 \cdot x_1(x_3+1) + 0.60022 \cdot x_2 \cdot (x_3+1) - 0.0097(x_3+1)^2 - 0.243x_3 \\ &+ 0.0394 \cdot x_1 \cdot x_3 - 0.60022 \cdot x_2 \cdot x_3 + 0.0097x_3^2 \\ &= 0.243 - 0.0394 \cdot x_1 + 0.60022 \cdot x_2 - 0.0097 \cdot 2 \cdot x_3 - 0.0097\end{aligned}$$

c) $R^2 = 1 - \frac{SCR}{SCT} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - \frac{209.55}{6777.5} = 0.969$

Ajuste bueno.

7) a) $\hat{\sigma}^2 = \dfrac{SCR}{n}, \tilde{\sigma}^2 = \dfrac{SCR}{n-k-1}$

$$\hat{\sigma}^2 = \frac{SCR}{n} = \frac{1950.7292}{6} = 325.0705$$
$$\tilde{\sigma}^2 = \frac{SCR}{n-k-1} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1} = 650.141$$
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - \frac{1950.423}{14112} = 1 - 0.1382 = \boxed{0.8618}$$

b) $\hat{y}_{x_1=25, x_2=1000} = 350.994 - 1.2799 \cdot 25 - 0.15390 \cdot 1000 = \boxed{165.29}$

8) En un proceso industrial se sospecha que la dureza de las láminas de acero reducido en frío depende del contenido en cobre (x_1) y de la temperatura de recocido (x_2). Para comprobar esta suposición se mide en doce especímenes de láminas de acero la dureza para varios valores del contenido de cobre y de la temperatura, obteniéndose los siguientes resultados:

Dureza (Rockwell 30-T)	Contenido de cobre (%) (x_1)	Temperatura de recocido (? F)(x_2)
79.1	0.025	1000
65.3	0.025	1100
55.5	0.025	1200
56.6	0.025	1300
81.1	0.15	1000
69.9	0.15	1100
57.6	0.15	1200
55.6	0.15	1300
85.5	0.2	1000
72.0	0.2	1100
60.9	0.2	1200
59.1	0.2	1300

a) Plantear el modelo de regresión lineal múltiple para explicar la dureza de la lámina en función del contenido de cobre y de la temperatura.

PLanteamiento general del modelo

$$Y = \theta' x + u = (\theta_0, \theta_1, \theta_2) \cdot \begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix} + u = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + u$$

$$y^{(i)} = \theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + u^{(i)}, \quad i = 1, \ldots, k$$

$$\underbrace{\begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} \\ \vdots & & \\ 1 & x_1^{(n)} & x_2^{(n)} \end{pmatrix}}_{M} \cdot \underbrace{\begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{pmatrix}}_{\theta} + \underbrace{\begin{pmatrix} u^{(1)} \\ u^{(2)} \\ \vdots \\ u^{(n)} \end{pmatrix}}_{U} \longrightarrow \begin{array}{l} Y = M \cdot \theta + U \\ U = Y - M \cdot \theta \end{array}$$

$$M'M = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(n)} \\ x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(n)} \end{pmatrix} \cdot \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} \\ \vdots & & \\ 1 & x_1^{(n)} & x_2^{(n)} \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^{n} x_1^{(i)} & \sum_{i=1}^{n} x_2^{(i)} \\ \sum_{i=1}^{n} x_1^{(i)} & \sum_{i=1}^{n} x_1^{(i)2} & \sum_{i=1}^{n} x_1^{(i)} x_2^{(i)} \\ \sum_{i=1}^{n} x_2^{(i)} & \sum_{i=1}^{n} x_1^{(i)} x_2^{(i)} & \sum_{i=1}^{n} x_2^{(i)2} \end{pmatrix} =$$

$$\begin{pmatrix} 12 & 1.5 & 13800 \\ 1.5 & 0.2525 & 1725 \\ 13800 & 1725 & 16020000 \end{pmatrix}$$

$$M'Y = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(n)} \\ x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(n)} \end{pmatrix} \cdot \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y^{(i)} \\ \sum_{i=1}^{n} x_1^{(i)} y^{(i)} \\ \sum_{i=1}^{n} x_2^{(i)} y^{(i)} \end{pmatrix} = \begin{pmatrix} 798.2 \\ 101.5425 \\ 905110 \end{pmatrix}$$

$$\hat{\theta} = (M'M)^{-1}M'Y = \begin{pmatrix} 9.14038 & -9230 & -0.076 \\ -1.9230 & 15.384 & 0 \\ -0.0076 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 798.2 \\ 101.5425 \\ 905110 \end{pmatrix} = \begin{pmatrix} 161.4042 \\ 27.1923 \\ -0.08549 \end{pmatrix}$$

b) Sabiendo que $\hat{\theta} = (161.404, 27.1923, -0.08549)'$, obtener una estimación para la varianza del error y el valor de R^2. Comentar la bondad del ajuste.

y	x_1	x_2	$\hat{y}$	$(y-\hat{y})^2$
79.1	0.025	1000	76.59	6.28
65.3	0.025	1100	68.05	7.54
55.5	0.025	1200	59.5	15.97
56.6	0.025	1300	50.95	31.96
81.1	0.15	1000	79.99	1.23
69.9	0.15	1100	71.44	2.38
57.6	0.15	1200	62.9	28.04
55.6	0.15	1300	54.35	1.57
85.5	0.2	1000	81.35	17.2
72.0	0.2	1100	72.8	0.65
60.9	0.2	1200	34.25	11.25
59.1	0.2	1300	55.71	11.52
				135.58

$$ECM = \frac{135.572}{12} = 11.298$$
$$\sigma^2 = \frac{135.572}{9} = 15.064$$
$$\sum_{i=1}^{n}(y^{(i)} - \overline{y})^2 = 1271.312$$
$$R^2 = 0.894$$

c) Se decide eliminar la variable x_1 del modelo y ajustar un modelo de regresión lineal simple para predecir la dureza en función de x_2.

1) ¿Qué modelo se obtendría?

$$y = \beta_0 + \beta_1 x_1 + U$$
$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x_2} \simeq = 165.159$$

2) ¿Qué cambio se espera en la dureza promedio si la temperatura se incrementa en 100 grados Farenheith?

$$\sum_{i=1}^{n} d^2(O_i, P) = \sum_{i=1}^{n}\sum_{j=1}^{n}(x_{ij} - P_j)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}(x_{ij} - \overline{x}_j + \overline{x}_j - P_j)^2$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\left((x_{ij} - \overline{x}_j)^2 + (\overline{x}_j - P_j)^2 + 2(x_{ij} - \overline{x}_j)(\overline{x}_j - P_j)\right)$$

$$= \sum_{i=1}^{n} d^2(O_i, \overline{O}) + \sum_{i=1}^{n}\sum_{j=1}^{n}(\overline{x}_j - P_j)^2 \geq \sum_{i=1}^{n} d^2(P_i, \overline{O})$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(x_{ij} - \overline{x}_j)(\overline{x}_j - P_j) = \sum_{j=1}^{n}(\overline{x}_j - P_j) \cdot \underbrace{\sum_{i=1}^{n}(x_{ij} - \overline{x_j})}_{\sum_{i=1}^{n} x_{ij} - n\overline{x}_j} = 0$$

0

$O_i = (x_{i,1}, \ldots, x_{i,k}) = (x_1^{(i)}, \ldots, x_k^{(i)})$

$P = (P_1, \ldots, P_k)$

$\overline{x}_j = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_j^{(i)}$

Tema 3: Regresión logística y multinomial

3.1) Modelo de regresión logística

3.1.1) Contexto

Deseamos predecir una variable binaria Y que solo toma los valores 0 y 1.

- Además, estos valores numéricos solo indicarán la pertenencia o no a un determinado grupo.

Ejemplos:

- Determinar si un paciente tiene o no una determinada enfermedad en función de diferentes variables (edad, presión arterial, nivel de colesterol, etc.).
  - En este caso el valor 1 suele indica que sí la tiene y 0 que no.
- Predecir si un estudiante aprueba o no un examen en función de las horas de estudio.
- Predecir si un mensaje de correo electrónico es spam o no en función de las palabras clave.
- Predecir si un cliente comprará o no un determinado producto en función de la edad y el salario.
- Predecir si un paciente tiene diabetes o no en función de variables como el nivel de glucosa, la presión arterial y el índice de masa corporal (IMC).

3.1.2) Objetivo

Objetivo: predecir la variable respuesta Y a partir de k variables numéricas $X_1, \ldots, X_k$ utilizando una única función

$$h_\theta(\mathbf{x}) = g(\theta'\mathbf{x}) = g(\theta_0 + \theta_1 x_1 + \cdots + \theta_k x_k),$$

donde $\theta = (\theta_0, \ldots, \theta, \theta_k)' \in \mathbb{R}^{k+1}$ contiene los parámetros del modelo y $\mathbf{X} = (X_0, \ldots, X_k)'$ las variables que podemos medir en nuestro individuos para predecir Y.

Para mejorar la notación hemos incluido una variable artificial X_0 siempre que vale 1.

3.1.3) ¿Cómo elegir la función g?

La función g debe transformar esos valores numéricos (lineales) en números entre 0 y 1 que nos indicarán la probabilidad de que el individuo pertenezca al grupo ($Y = 1$):

$$g : \mathbb{R} \to [0, 1]$$

y $h_\theta(\mathbf{x}) \approx Pr(Y = 1|\mathbf{X} = \mathbf{x})$, donde $\mathbf{X} = (X_0, \ldots, X_k)'$.

Regla de decisión:

$$h_\theta(\mathbf{x}) \geq 0.5 \to \hat{y} = 1$$
$$h_\theta(\mathbf{x}) < 0.5 \to \hat{y} = 0$$

donde $\hat{y}$ representa el valor que predecimos para Y cuando $\mathbf{X} = \mathbf{x}$.

Existen diversas opciones para determinar g, la más popular es la función logística (o sigmoide)

$$g(z) = \frac{1}{1 + \exp(-z)} = \frac{\exp(z)}{1 + \exp(z)}.$$

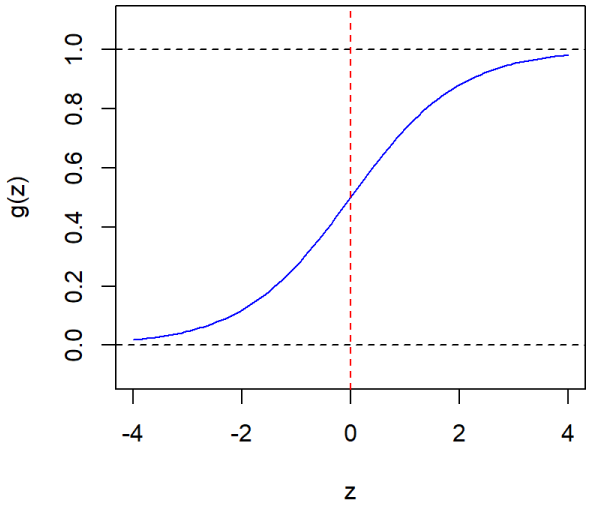
3.2) Función logística

```r
par(mfrow = c(1, 1))
g <- function(x) 1/(1+exp(-x))
x <- seq(-4, 4, length = 100)
y <- g(x)
plot(x, y, xlab = 'z', ylab = 'g(z)', col
    = 'blue', type = "l", ylim = c(-0.1,
    1.1), yaxp = c(0, 1, 5))
abline(h = 0, lty = 2)
abline(h = 1, lty = 2)
abline(v = 0, lty = 2, col = "red")
```

- Propiedades
  - Es continua
  - Estrictamente creciente.
  - Recorrido de 0 a 1.
  - Transformará el valor $\theta'\mathbf{x} \in \mathbb{R}$ en un valor $h_\theta(\mathbf{x}) \in [0, 1]$.
  - $g(0) = 0.5$
  - Regla de decisión:

$$\theta'\mathbf{x} \geq 0 \to \hat{y} = 1$$
$$\theta'\mathbf{x} < 0 \to \hat{y} = 0$$

  - La función $I_\theta(\mathbf{x}) = \theta'\mathbf{x}$ define un índice de separación entre las categorías de Y



3.3) ¿Cómo determinar una función costo que penalice las decisiones erróneas?

3.3.1) Función costo

Para $z = h_\theta(\mathbf{x})$, definimos la función costo

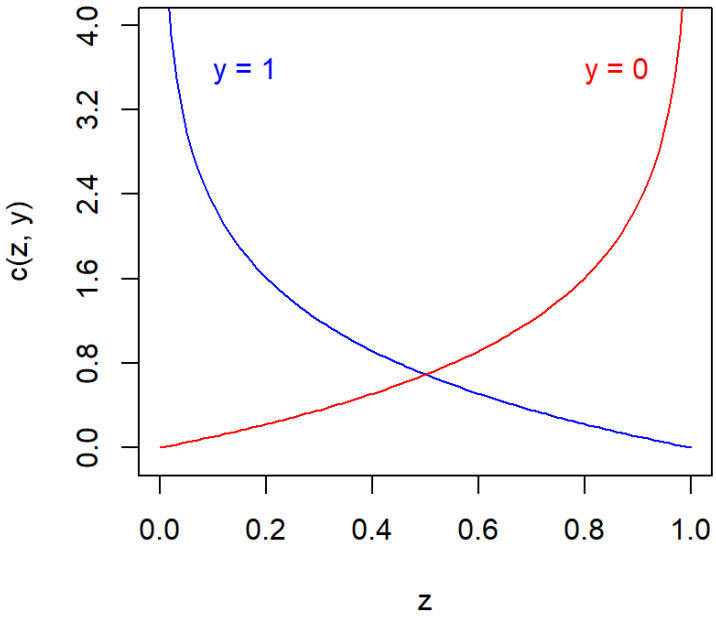
$$c(z, y) = \begin{cases} -\log(z) & \text{si } y = 1 \\ -\log(1 - z) & \text{si } y = 0 \end{cases}$$

O, equivalentemente,

$$c(z, y) = -y \log(z) - (1 - y) \log(1 - z),$$

donde $y \in \{0, 1\}$ y los logaritmos son neperianos.

```r
par(mfrow = c(1, 1))
x <- seq(0, 1, length = 100)
plot(x, -log(x), xlab = 'z', ylab = 'c(z, y)',
    col = 'blue', type = "l",  ylim = c(-0.1, 4)
    , yaxp = c(0, 4, 5))
text(0.1, 3.5, "y = 1", col = "blue", adj = c
    (0, 0))
lines(x, -log(1-x), xlab = 'z', ylab = 'c(z, y)
    ', col = 'red', type = "l",  ylim = c(-0.1,
    4), yaxp = c(0, 4, 5))
text(0.8, 3.5, "y = 0", col = "red", adj = c(0,
     0))
```



3.3.2) Criterio

Minimizar el valor esperado de la función costo

$$\min_{\theta} J(\theta) = E[c(h_\theta(\mathbf{X}), Y)]$$

Para determinar los valores óptimos de los parámetros:

- Dispondremos de una muestra (training sample) y de individuos en los que se conozcan tanto los valores de $\mathbf{x}$ como los valores de y (aprendizaje supervisado).
- Calcularemos los costos en los valores muestrales.
- Determinaremos los valores de los parámetros que minimizan estos costos.

3.3.3) Otra formulación del problema

Si $p = Pr(Y = 1)$ este modelo es equivalente a suponer que existe una relación lineal entre las variables $X_1, \ldots, X_k$ y la función `log-odd de p`.

$$\log \frac{p}{1-p} = \theta' \mathbf{X}.$$

Puesto que esto es equivalente a suponer que

$$p = Pr(Y = 1) = \frac{\exp(\theta' \mathbf{X})}{1 + \exp(\theta' \mathbf{X})} = g(\theta' \mathbf{X})$$

3.4) Inferencia y predicción

3.4.1) Función costo empírica

Datos muestrales: $(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(n)}, y^{(n)})$.

Función costo:

$$J(\theta) \coloneqq \frac{1}{n}\sum_{i=1}^{n} c(h_\theta(\mathbf{x}^{(i)}), y^{(i)}).$$

Desarrollando la función c obtenemos

$$\begin{aligned} J(\theta) &= -\frac{1}{n}\sum_{i=1}^{n}\left[y^{(i)}\log(h_\theta(\mathbf{x}^{(i)})) + (1-y^{(i)})\log(1-h_\theta(\mathbf{x}^{(i)}))\right] \\ &= -\frac{1}{n}\sum_{i=1}^{n}\left[y^{(i)}\log(g(\theta'\mathbf{x}^{(i)})) + (1-y^{(i)})\log(1-g(\theta'\mathbf{x}^{(i)}))\right] \end{aligned}$$

3.4.2) Función costo empírica en forma matricial

Denotando

- $M = (x_j^{(i)})$ a la matriz de datos.
- $\mathbf{y} = (y^{(i)})$ al vector columna con los valores de Y
- $h \coloneqq g(M\theta)$ al vector columna con los ajustes en cada individuo, entonces

$$J(\theta) \coloneqq -\frac{1}{n}\left[\mathbf{y}'\log(h) + (1_n - \mathbf{y})'\log(1_n - h)\right]$$

  donde 1_n representa un vector columna de dimensión n.

3.4.3) Objetivo

Ajsutar el parámetro θ para que J tome el menor valor posible.

- Solución: Algoritmos iterativos de búsqueda como, por ejemplo, el algoritmo del gradiente descendente.
  - Práctica complementaria de regresión logística.
- Existen varias librerías de R que permiten obtener estimaciones de los parámetros del modelo logístico.
  - Práctica de regresión logística.

3.5) Un ejemplo sencillo

3.5.1) Datos muestrales

Como en técnicas anteriores usaremos un ejemplo sencillo para comprobar cómo funciona nuestro modelo.

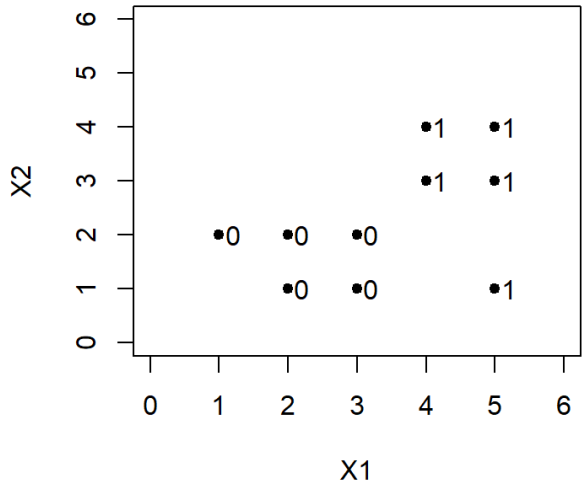
Supongamos que tenemos dos variables predictoras X_1 y X_2 ($k = 2$) y los datos siguientes:

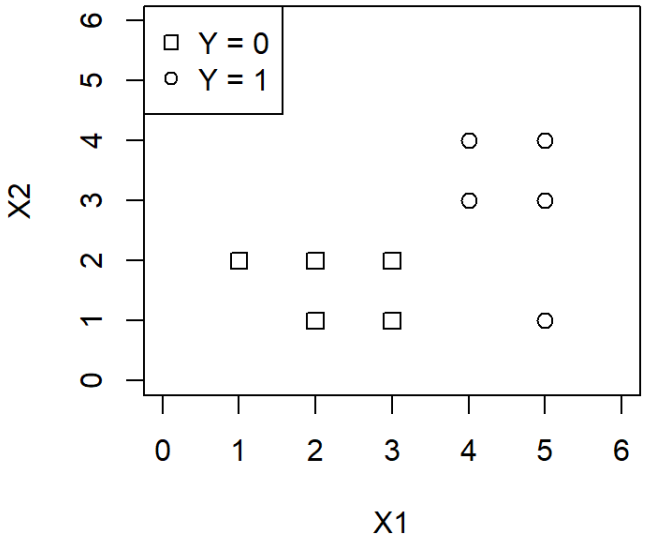
Individuo	X_1	X_2	Y
1	1	2	0
2	2	1	0
3	3	1	0
4	2	2	0
5	5	1	1
6	5	3	1
7	3	2	0
8	4	3	1
9	4	4	1
10	5	4	1

Lo primero que tenemos que hacer (si es posible) es dibujar estos puntos añadiendo una etiqueta para distinguir los de cada grupo.

```
X1 <- c(1, 2, 3, 2, 5, 5, 3, 4, 4, 5)
X2 <- c(2, 1, 1, 2, 1, 3, 2, 3, 4, 4)
Y <- c(0, 0, 0, 0, 1, 1, 0, 1, 1, 1)
plot(X1, X2, xlab = "X1", ylab = "X2",
     pch = 20, xlim = c(0,6), ylim = c
    (0,6), cex = 1.2)
text(X1 + 0.2, X2, Y, cex = 1)
```

```
X1 <- c(1, 2, 3, 2, 5, 5, 3, 4, 4, 5)
X2 <- c(2, 1, 1, 2, 1, 3, 2, 3, 4, 4)
Y <- c(0, 0, 0, 0, 1, 1, 0, 1, 1, 1)
plot(X1, X2, xlab = "X1", ylab = "X2", pch
     = as.integer(Y), xlim = c(0,6), ylim =
     c(0,6), cex = 1.2)
legend('topleft', legend = c('Y = 0','Y =
    1'), pch = 0:1, cex = 1)
```





En ambas gráficas podemos observar que los dos grupos se pueden separar muy bien con rectas.

Por lo tanto nuestro modelo será

$$h_\theta(\mathbf{x}) = g(\theta_1 + \theta_1 x_1 + \theta_2 x_2).$$

Otra forma de analizar los grupos es calcular medidas descriptivas en cada uno de ellos.

- Por ejemplo podemos calcular las medias en cada grupo:

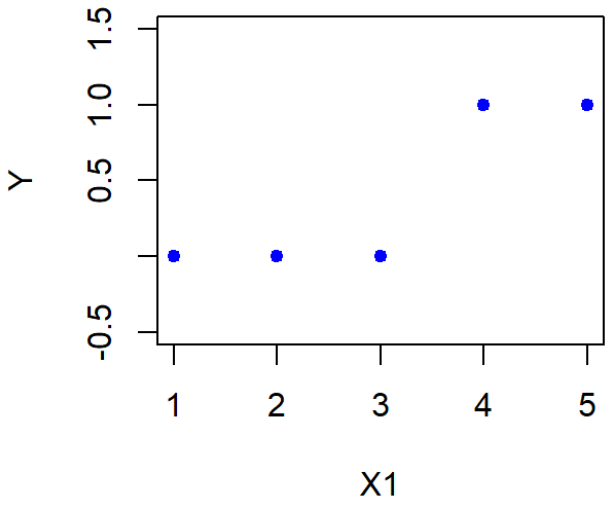
```r
X1 <- c(1, 2, 3, 2, 5, 5, 3, 4, 4, 5)
X2 <- c(2, 1, 1, 2, 1, 3, 2, 3, 4, 4)
Y <- c(0, 0, 0, 0, 1, 1, 0, 1, 1, 1)
tmp1 = tapply(X1, Y, mean)
tmp2 = tapply(X2, Y, mean)
tmp = rbind(tmp1, tmp2)
colnames(tmp) = paste0("Y = ", colnames(tmp))
rownames(tmp) = paste0("Media de ", c("X1", "X2"))
tmp
```

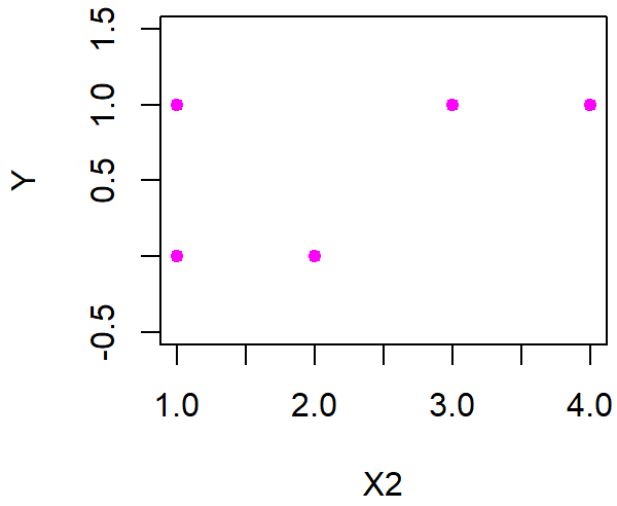
```
##             Y = 0 Y = 1
## Media de X1   2.2   4.6
## Media de X2   1.6   3.0
```

Estas diferencias también se pueden ver representado $x^{(i)}$ frente a Y.

```r
X1 <- c(1, 2, 3, 2, 5, 5, 3, 4, 4, 5)
Y <- c(0, 0, 0, 0, 1, 1, 0, 1, 1, 1)
plot(X1, Y, ylim = c(-0.5, 1.5), cex =
    1.2, pch = 20, col = "blue")
```

```r
X2 <- c(2, 1, 1, 2, 1, 3, 2, 3, 4, 4)
Y <- c(0, 0, 0, 0, 1, 1, 0, 1, 1, 1)
plot(X2, Y, ylim = c(-0.5, 1.5), cex =
    1.2, pch = 20, col = "magenta")
```





Podemos observar cómo la primera variable separa mejor a los grupos que la segunda (en los valores muestrales).

De forma similar se pueden representar histogramas o diagramas caja-bigote para comparar las variables en cada grupo.

Podemos calcular la función $J(\theta)$ en R con los datos anteriores.

```r
X1 <- c(1, 2, 3, 2, 5, 5, 3, 4, 4, 5)
X2 <- c(2, 1, 1, 2, 1, 3, 2, 3, 4, 4)
Y <- c(0, 0, 0, 0, 1, 1, 0, 1, 1, 1)
n <- length(Y)
k <- 2
X0 = rep(1, n)
M <- matrix(c(X0, X1, X2), nrow = n, ncol = k + 1, byrow = FALSE)
g <- function(z){
  g = exp(z)/(1 + exp(z))
  return(g)
}
J <- function(theta){
  J = - sum(Y * log(g(M %*% theta)) + (1 - Y) * log(1 - g(M %*% theta)))/n
  return(J)
}
```

Podemos aplicar un método iterativo para la obtención del óptimo.

- Por ejemplo, el método del gradiente descendiente (se detallará su aplicación en la práctica complementaria de regresión logística).

```r
z <- c(-3.5, 1, 0)
alpha <- 1/3
m <- 1000
J1 <- 1:m
for (i in 1:m) {
  h <- g(M %*% z)
  z <- z - (alpha/n) * t(M) %*% (h-Y)
  J1[i] <- J(z)
}
```

Partiendo del punto inicial $\theta^{(0)} = (-3.5,\ 1,\ 0)$ (recta vertical de separación $x_1 = 3.5$) y después de 1000 iteraciones obtendremos $\hat{\theta}_0 = -10.7505$, $\hat{\theta}_1 = 2.4594$ y $\hat{\theta}_3 = 1.0287$, con valor de $J(\hat{\theta}) = 0.0597$.

En este caso,

$$I_{\hat{\theta}}(x_{1,2}) = -10.7505 + 2.4591x_1 + 1.0287x_2$$

La recta que marca la frontera de esta solución será

$$-10.7505 + 2.4594x_1 + 1.0284x_2 = 0$$

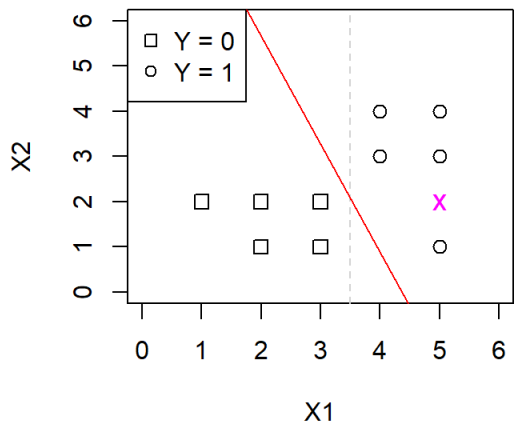
Esto es, $x_2 = 10,4506 - 2.3908x_1$

Si queremos predecir el grupo para un nuevo individuo con valores $x_1 = 5$ y $x_2 = 2$

- Evaluamos la función $I_{\hat{\theta}}(x_{1,2}) = -10.7505 + 2.4594x_1 + 1.0287x_2$
- $I_{\hat{\theta}}(5, 2) = 3.6037 > 0$, por lo que el individuo se clasifica en el grupo $y = 1$.

Representamos los valores muestrales incluyendo la recta que marca la frontera y el punto $(5, 2)$.

```
plot(X1, X2, xlab = "X1", ylab = "X2", pch = as.integer(Y), xlim = c(0,6), ylim = c
    (0,6), cex = 1.2)
legend('topleft', legend = c('Y = 0','Y = 1'), pch = 0:1, cex = 1)
abline(v = 3.5, col = "lightgray", lty = 2)
abline(-z[1]/z[3],-z[2]/z[3],col='red')
text(5, 2, 'x', col = "magenta", cex = 1.2)
```



Para medir cómo de fiable es esta clasificación podemos:

- Observar cómo se distribuyen los puntos en esta gráfica (cuando sea posible).
- Calcular las probabilidades a posteriori.

$Pr(Y = 1|X_1 = 5, X_2 = 2) \approx g(I_{\hat{\theta}}(5, 2)) = 0.9735$

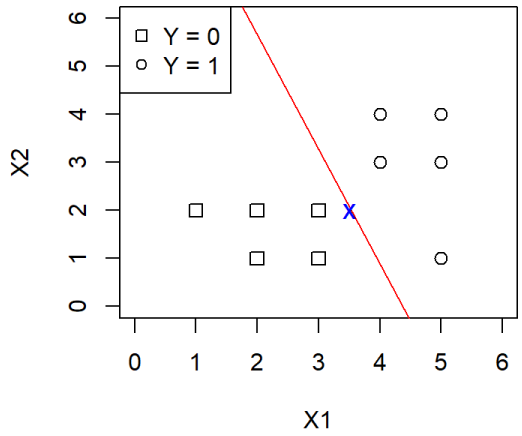
$Pr(Y = 0|X_1 = 5, X_2 = 2) \approx 1 - g(I_{\hat{\theta}}(5, 2)) = 0.0265$

Observando la gráfica con los valores muestrales parece razonable que el nuevo individuo con valores $\mathbf{x} = (5, 2)$ sea clasificado en el grupo $y = 1$.

Ahora queremos predecir el grupo para otro nuevo individuo con valores $x_1 = 3.5$ y $x_2 = 2$. En este caso: $I_{\hat{\theta}}(x_1, x_2) = -0.0853 < 0$, por lo que el individuo se clasifica en el grupo $y = 0$.

Representamos gráficamente:

```
plot(X1, X2, xlab = "X1", ylab = "X2", pch = as.integer(Y), xlim = c(0,6), ylim = c
    (0,6), cex = 1.2)
legend('topleft', legend = c('Y = 0','Y = 1'), pch = 0:1, cex = 1)
abline(-z[1]/z[3],-z[2]/z[3],col='red')
text(3.5, 2, 'x', col = "blue", cex = 1.2)
```



Observamos dónde se encuentra el nuevo individuo en la gráfica.

Calculamos las probabilidades a posteriori:

- $Pr(Y = 1|X_1 = 3.5, X_2 = 2) \approx g(I_{\hat{\theta}}(3.5, 2)) = 0.4787$
- $Pr(Y = 0|X_1 = 3.5, X_2 = 2) \approx 1 - g(I_{\hat{\theta}}(3.5, 2)) = 0.5213$

Recordemos que en realidad no estamos seguros de que esos valores sean realmente esas probabilidades.

De esta forma intuimos que esta clasificación no es muy fiable ya que ese punto está muy cerca de la frontera.

3.6) Regresión logística multinomial

3.6.1) Contexto

Generalización del modelo de regresión logística binaria.
La variable dependiente tiene más de dos categorías, sin/con un orden implícito.

- Primer caso: considera variables de respuesta nominal,
  - Por ejemplo, el país de procedencia, el color de un automóvil, etc.
- Segundo caso: trata variables de respuesta ordinal,
  - Por ejemplo, el nivel educativo, la fase de una enfermedad, etc.

3.6.2) Objetivo

Estimar la probabilidad de que un individuo presente cada una de estas categorías en función de los valores que se observen de las variables explicativas.

3.7) Modelo teórico

3.7.1) Formulación

La variable respuesta Y puede presentar g características.

- Y toma los valores $1, 2, \ldots, g$, que indican la pertenencia a cada grupo definido por cada categoría, con probabilidades $p_1, p_2, \cdots, p_g$, respectivamente, tales que
$$\sum_{j=1}^{g} p_j = 1.$$
Consideramos como referencia una de las categorías, por ejemplo, la última, g.

Establecemos un modelo `logit` para cada categoría con respecto a esta:

$$\log \frac{p_j}{p_g} = \log \frac{Pr[Y=j]}{Pr[Y=g]} = \theta_{\mathbf{j}}' \mathbf{X}, \quad j = 1, \ldots, g-1,$$

donde $\theta_j = (\theta_{j,0}, \ldots, \theta_{hk})' \in \mathbb{R}^{k+1}$ contiene los parámetros del modelo para cada categoría j y $\mathbf{X} = (X_0, \ldots, X_k)'$ las variables que podemos medir en nuestros individuos para predecir Y.

Al cociente $\dfrac{p_j}{p_g}$ se le denomina `odds` de la categoría j respecto de la categoría g.

Se ha considerado un término constante en el modelo incluyendo la variable artificial X_0 que siempre vale 1.

Cada uno de los coeficientes se interpreta como el efecto de cada variable explicativa sobre el logaritmo de los `odds` de la categoría j respecto de la categoría de referencia g.

Cuando $g = 2$, el modelo se reduce a una única ecuación equivalente a la propuesta en la regresión logística.

3.7.2) Observaciones

Si comparamos las probabilidades para dos categorías diferentes, i y j, utilizando el modelo anterior obtenemos que:

$$\log \frac{p_i}{p_j} = \log \frac{\frac{p_i}{p_g}}{\frac{p_j}{p_g}} = \log \frac{p_i}{p_g} - \log \frac{p_j}{p_g} = \theta_{\mathbf{i}}' \mathbf{X} - \theta_{\mathbf{j}}' \mathbf{X}$$
$$= (\theta_i - \theta_j)' \mathbf{X} = (\theta_{i0} - \theta_{j0}) + (\theta_{i1} - \theta_{j1}) X_1 + \cdots + (\theta_{ik} - \theta_{jk}) X_k.$$

De esta forma, se obtiene una ecuación `logit` de la categoría i con respecto a la categoría j, donde $\theta_0 = \theta_{i0} - \theta_{j0}$, $\theta_1 = \theta_{i1} - \theta_{j1}, \ldots, \theta_k = \theta_{ik} - \theta_{jk}$.

3.7.2.1) Un ejemplo ficticio

Supongamos que deseamos estudiar cómo influye el sexo del neonato en la aparición de determinados problemas durante el parto.

Se contemplan únicamente tres opciones posibles para los partos (Y):

- $Y = 1$: parto con el problema A
- $Y = 2$: parto con el problema B
- $Y = 3$: parto sin problemas

La tercera opción se toma como la opción de referencia.

Se introduce una variable binaria X para representar el sexo del neonato:

- $X = 0$: si es niño
- $X = 1$: si es niña

Planteamos un modelo de regresión logística para predecir la variable categórica Y en función de X:

$$\log \frac{p_j}{p_3} = \log \frac{Pr[Y=j]}{Pr[Y=3]} = \theta_{j0} + \theta_{j1}x, \quad j = 1, 2.$$

3.7.2.2) Interpretación de los parámetros

Considerando la primera de las ecuaciones:

$$\log \frac{p_1}{p_3} = \log \frac{Pr[Y=1]}{Pr[Y=3]} = \theta_{10} + \theta_{11}x$$

- Si $X = 0$, $\dfrac{p_1}{p_3} = \exp(\theta_{10})$.
- Si $X = 1$, $\dfrac{p_1}{p_3} = \exp(\theta_{10} + \theta_{11})$.

Por lo tanto, respecto a la probabilidad de un parto normal la probabilidad de la presencia del problema A se multiplica por $\exp(\theta_{10} + \theta_{11})$ en el caso de niños y por $\exp(\theta_{10})$ en el caso de niñas.

Si, por ejemplo, resultara $\theta_{11} = 0$, el sexo no influiría sobre la probabilidad de que aparezca el problema A.

Razonando de forma análoga, si resultara $\theta_{21} > 0$ se concluiría que la aparición del problema B es más probable en niñas que en niños.

3.7.2.3) Estimador de máxima verosimilitud

Sea X una variable aleatoria, con función de densidad o función puntual de probabilidad $x \longmapsto f_X(x, \theta)$, donde $\theta \in \Theta \subset \mathbb{R}^k$.

Consideramos una muestra aleatoria simple $(X_1, \ldots, X_n)$.

Para un valor concreto de $(X_1, \ldots, X_n)$, que denotamos por $(x_1, \ldots, x_n)$, la función de verosimilitud L_n es una función de θ, $L_n : \Theta \subset \mathbb{R}^k \longrightarrow \mathbb{R}^+$, definida como

$$L_n(\theta) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f_{\mathbf{X}}(x_i; \theta).$$

El estimador de máxima verosimilitud $\hat{\theta}$ de θ es cualquier valor de θ admisible que maximiza la función $L_n(\theta)$,

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L_n(\theta).$$

3.8) Criterio de máxima de verosimilitud para nuestro modelo

3.8.1) Criterio

Para estimar los parámetros del modelo utilizaremos el criterio de máxima verosimilitud:

- Calculamos la función de verosimilitud.
- Maximizamos esta función para obtener los estimadores de máxima verosimilitud (MLE, `maximum likelihood estimator`).

Redefiniremos la variable Y en g variables indicadoras $(Y_1, \ldots, Y_g)$:

- Y_j toma el valor 1 si la respuesta pertenece al grupo j y 0 en otro caso.
- Tendremos que $\sum_{j=1}^{g} Y_j = 1$.

3.8.2) Función de verosimilitud

Supongamos que disponemos de n observaciones independientes de la variable Y y de las variables explicativas.Para cada individuo i tendremos:

- Las observaciones $(y_1^{(i)}, \ldots, y_g^{(i)})$, donde
$$\sum_{j=1}^{g} y_j^{(i)} = 1.$$
- Los valores de las variables explicativas observados $\mathbf{x}^{(i)} = (x_0^{(i)}, \ldots, x_k^{(i)})'$.

Entonces, la función de verosimilitud adopta la expresión

$$L(\theta_1, \ldots, \theta_{g-1}) \propto \prod_{i=1}^{n} \prod_{j=1}^{g} p_j(\mathbf{x}^{(i)})^{y_j^{(i)}}$$

donde el símbolo $\propto$ indica proporcionalidad.

3.8.3) Función de log-verosimilitud

En lugar de maximizar directamente la función de verosimilitud consideraremos su logaritmo neperiano:

- Función más manejable que simplifica los cálculos.
- Permite utilizar métodos numéricos de optimización más eficientes y estables al transformar productos en sumas.

La log-verosimilitud adopta la expresión

$$\log L(\theta_1, \ldots, \theta_{g-1}) \propto \log \prod_{i=1}^{n} \prod_{j=1}^{g} p_j(\mathbf{x}^{(i)})^{y_k^{(i)}} = \sum_{i=1}^{g} \sum_{j=1}^{g} y_j^{(i)} \log p_j(\mathbf{x}^{(i)}).$$

En términos de una función costo, el criterio de máxima verosimilitud equivale a minimizar la función

$$J(\theta_1, \ldots, \theta_{g-1}) = -\log L(\theta_1, \ldots, \theta_{g-1}).$$

En ocasiones en lugar de la función de log-verosimilitud se utiliza la función auxiliar $\Lambda = -2 \log L$, denominada la `deviance` del modelo.

Puesto que $p_g = 1 - (p_1 + \cdots + p_{g-1})$, la contribución del individuo i en la función de la log-verosimilitud sería

$$\begin{aligned}\sum_{j=1}^{g} y_j^{(i)} \log p_j(\mathbf{x}^{(i)}) &= \sum_{j=1}^{g-1} y_j^{(i)} \log p_j(\mathbf{x}^{(i)}) + \left(1 - \sum_{j=1}^{g-1} y_j^{(i)}\right) \log \left(1 - \sum_{j=1}^{g-1} p_j(\mathbf{x}^{(i)})\right) \\ &= \sum_{j=1}^{g-1} y_j^{(i)} \log \frac{p_j(\mathbf{x}^{(i)})}{1 - \sum_{h=1}^{g-1} p_h(\mathbf{x}^{(i)})} + \log \left(1 - \sum_{j=1}^{g-1} p_j(\mathbf{x}^{(i)})\right).\end{aligned}$$

Según el modelo logístico multinomial,

$$\log \frac{p_j(\mathbf{x}^{(i)})}{p_g(\mathbf{x}^{(i)})} = \boldsymbol{\theta_j}^{(i)}.$$

Por otra parte, $p_g(\mathbf{x}^{(i)})$ puede escribirse como

$$p_g(\mathbf{x}^{(i)}) = 1 - \sum_{j=1}^{g-1} p_j(\mathbf{x}^{(i)}) = 1 - p_g(\mathbf{x}^{(i)}) \sum_{j=1}^{g-1} \exp(\boldsymbol{\theta}_\mathbf{j}' \mathbf{x}^{(i)}).$$

Y despejando ahora $p_g(\mathbf{x}^{(i)})$ en la expresión anterior se obtiene que

$$p_g(\mathbf{x}^{(i)}) = \frac{1}{1 + \sum_{h=1}^{g-1} \exp(\theta_j' \mathbf{x}^{(i)})}.$$

Y por tanto,

$$p_j(\mathbf{x}^{(i)}) = \frac{\exp(\theta_j' \mathbf{x}^{(i)})}{1 + \sum_{h=1}^{g-1} \exp(\theta_h' \mathbf{x}^{(i)})}.$$

Sustituyendo ahora en la función de log-verosimilitud se tendrá que

$$\begin{aligned}\log L(\theta_1, \ldots, \theta_{g-1}) \propto \sum_{i=1}^{g}\sum_{j=1}^{g} y_j^{(i)} \log p_j(\mathbf{x}^{(i)}) &= \sum_{i=1}^{n}\left[\sum_{j=1}^{n} y_j^{(i)}(\theta_j' \mathbf{x}^{(i)}) - \log\left(1 + \sum_{j=1}^{g-1}\exp(\theta_j' \mathbf{x}^{(i)})\right)\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{g-1} y_j^{(i)}(\theta_j' \mathbf{x}^{(i)}) - \sum_{i=1}^{g-1}\log\left(1 + \sum_{j=1}^{g-1}\exp(\theta_j' \mathbf{x}^{(i)})\right)\end{aligned}$$

3.8.4) ¿Cómo obtenemos en la práctica las estimaciones de $\theta_1, \ldots, \theta_{g-1}$?

Para obtener valores de los parámetros $\theta_1, \ldots, \theta_{g-1}$ que maximicen la log-verosimilitud (o equivalentemente, minimicen la función costo J), podremos:

- Aplicar algoritmos iterativos de búsqueda como el algoritmo del gradiente descendente.
- Hacer uso de funciones implementadas en librerías de `R` que permiten obtener estimaciones de los parámetros del modelo logístico multinomial, como la función `multinom()` de la librería `nnet`.
  - Práctica de regresión logística multinomial.

3.9) Un caso sencillo

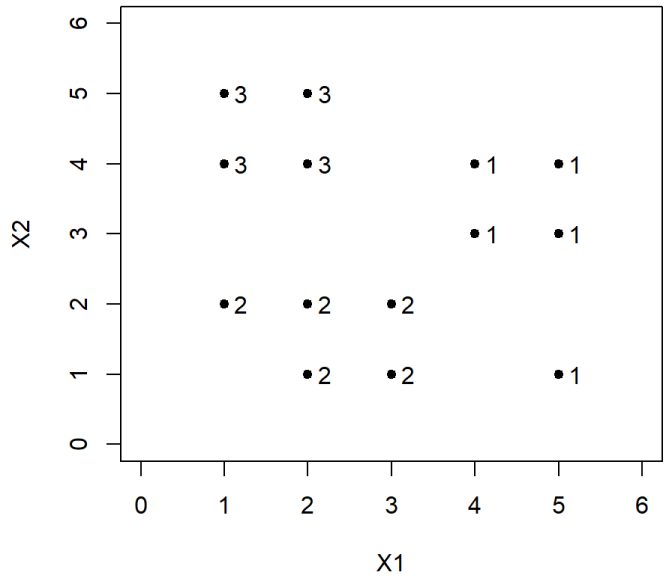
3.9.1) Cálculo de la verosimilitud

Supongamos que tenemos una variable respuesta (Y) con tres categorías posibles y dos variables explicativas X_1 y X_2, cuyas observaciones están recogidas en la tabla adjunta.

Individuo	X_1	X_2	Y
1	1	2	2
2	2	1	2
3	1	5	3
4	2	4	3
5	3	1	2
6	2	2	2
7	2	5	3
8	5	1	1
9	5	3	1
10	3	2	2
11	4	3	1
12	4	4	1
13	5	4	1
14	1	4	3

Para la implementación en R de estos cálculos introduciremos los datos en vectores y representaremos los datos gráficamente.

```
X1 <- c(1, 2, 1, 2, 3, 2, 2, 5, 5, 3, 4, 4, 5, 1)
X2 <- c(2, 1, 5, 4, 1, 2, 5, 1, 3, 2, 3, 4, 4, 4)
Y  <- c(2, 2, 3, 3, 2, 2, 3, 1, 1, 2, 1, 1, 1, 3)
plot(X1, X2, xlab = "X1", ylab = "X2", pch = 20, xlim = c(0,6), ylim = c(0,6), cex =
    1.2)
text(X1 + 0.2, X2, Y, cex = 1)
```



Incluimos los valores de la variable artificial $X_0 = 1$.

Asociada a la variable Y definimos tres variables indicadoras, Y_j, $j = 1, 2, 3$, de manera que $Y_j = 1$ si $Y = j$ y 0 en otro caso.

```
library("dplyr")
X1 <- c(1, 2, 1, 2, 3, 2, 2, 5, 5, 3, 4, 4, 5, 1)
X2 <- c(2, 1, 5, 4, 1, 2, 5, 1, 3, 2, 3, 4, 4, 4)
Y  <- c(2, 2, 3, 3, 2, 2, 3, 1, 1, 2, 1, 1, 1, 3)
df = data.frame(X1, X2, Y)
df <- df %>%
  mutate(X0 = 1,
         Y1 = ifelse(Y == 1, 1, 0),
         Y2 = ifelse(Y == 2, 1, 0),
         Y3 = ifelse(Y == 3, 1, 0))
df
```

```
##    X1 X2 Y X0 Y1 Y2 Y3
## 1   1  2 2  1  0  1  0
## 2   2  1 2  1  0  1  0
## 3   1  5 3  1  0  0  1
## 4   2  4 3  1  0  0  1
## 5   3  1 2  1  0  1  0
## 6   2  2 2  1  0  1  0
```

```
## 7   2  5 3  1  0  0  1
## 8   5  1 1  1  1  0  0
## 9   5  3 1  1  1  0  0
## 10  3  2 2  1  0  1  0
## 11  4  3 1  1  1  0  0
## 12  4  4 1  1  1  0  0
## 13  5  4 1  1  1  0  0
## 14  1  4 3  1  0  0  1
```

Empezamos implementando la función J utilizando código en R.

```
J <- function(theta){
  C = exp(t(theta)%*%t(X))
  D = colSums(C)
  E = matrix(rep(1 + D, g - 1), nrow = g - 1, ncol = n, byrow = TRUE)
  P = C/E
  Pg = 1/(1+D)
  PT =  rbind(P, Pg)
  J = -sum(YY*log(t(PT)))
  return(J)
}
```

Introducimos los valores muestrales de las variables.

```
library("dplyr")
n = length(Y)
g = 3
k = 2
X1 <-c(1, 2, 1, 2, 3, 2, 2, 5, 5, 3, 4, 4, 5, 1)
X2 <-c(2, 1, 5, 4, 1, 2, 5, 1, 3, 2, 3, 4, 4, 4)
Y  <-c(2, 2, 3, 3, 2, 2, 3, 1, 1, 2, 1, 1, 1, 3)
df = data.frame(X1, X2, Y)
df <- df %>%
  mutate(X0 = 1,
         Y1 = ifelse(Y == 1, 1, 0),
         Y2 = ifelse(Y == 2, 1, 0),
         Y3 = ifelse(Y == 3, 1, 0))
X = as.matrix(df[, c("X0", "X1", "X2")])
YY = as.matrix(df[, c("Y1", "Y2", "Y3")])
```

Evaluamos en

$$\theta = \begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 2 & 3 \end{pmatrix}$$

```
theta = matrix(c(1, 2,
                 0, 1,
                 2, 3), nrow = k+1, ncol = g-1, byrow = TRUE)
J(theta)
```

```
## [1] 111.0598
```

Evaluamos en

$$\theta = \begin{pmatrix} -2.3 & 1.5 \\ 0.5 & 2 \\ 2.1 & 3.5 \end{pmatrix}$$

```
theta = matrix(c(-2.3, 1.5,
                 0.5, 2,
                 2.1, 3.5), nrow = k+1, ncol = g-1, byrow = TRUE)
J(theta)
```

```
## [1] 155.5009
```

3.9.2) Estimación de los parámetros

Utilizaremos la función `multinom()` de la librería `nnet` de `R`.

- Las opciones de esta función se verán con más detalle en la práctica de regresión logística multinomial.

```
library("nnet")
X1 <-c(1, 2, 1, 2, 3, 2, 2, 5, 5, 3, 4, 4, 5, 1)
X2 <-c(2, 1, 5, 4, 1, 2, 5, 1, 3, 2, 3, 4, 4, 4)
Y  <-c(2, 2, 3, 3, 2, 2, 3, 1, 1, 2, 1, 1, 1, 3)
df = data.frame(X1, X2, Y)
df$Y_factor = factor(df$Y, levels = c("1", "2", "3"))
df$Y_factor <- relevel(df$Y_factor, ref = "3")
mymultinom <- multinom(Y_factor ~ X1 +  X2, data = df)
```

```
## # weights:  12 (6 variable)
## initial  value 15.380572
## iter  10 value 0.194988
## iter  20 value 0.010826
## iter  30 value 0.006357
## iter  40 value 0.005358
## iter  50 value 0.004594
## iter  60 value 0.003156
## iter  70 value 0.002732
## iter  80 value 0.002396
## iter  90 value 0.002072
## iter 100 value 0.001924
## final  value 0.001924
## stopped after 100 iterations
```

```
summary(mymultinom)$coefficients
```

```
##   (Intercept)          X1           X2
## 1 -19.79766347 12.677891791  -5.560148528
## 2  27.10761881  2.521385121 -10.282168169
```

3.9.3) Modelo estimado

Ecuaciones del modelo estimado:

$$\log \frac{p_1}{p_3} = -19.798 + 12.678x_1 - 5.56x_2$$
$$\log \frac{p_2}{p_3} = 27.108 + 2.521x_1 - 10.282x_2$$

Recordemos que los coeficientes del modelo miden la variación del logaritmo de los `odds` por unidad de cambio en el correspondiente predictor.

Tomando exponenciales sobre los coeficientes, medimos las variaciones producidas sobre los `odds` directamente.

El algoritmo ha parado después de 100 iteraciones.

Valor de la $-\log L$: 0.0019242

Valor de la `deviance` del modelo: 0.0038484

3.10) Modelo logístico multinomial con categorías ordinales

3.10.1) Modelo teórico

Este tipo de modelo se utiliza cuando las categorías de la variable dependiente representan un orden lógico o jerarquía.

Expresión general del modelo:

$$\log \frac{Pr[Y \leq j]}{1 - Pr[Y \leq j]} = \theta_j' \mathbf{X}, \quad j = 1, \ldots, g-1,$$

donde $\theta_j = (\theta_{j0}, \ldots, \theta_{jk})' \in \mathbb{R}^{k+1}$ contiene los parámetros del modelo para cada categoría j y $\mathbf{X} = (X_0, \ldots, X_k)'$ las variables que podemos medir en nuestros individuos para predecir Y.

Interpretación de los coeficientes: entender cómo un cambio en una variable predictora afecta a la razón de probabilidades de que la variable dependiente sea menor o igual a una categoría específica en comparación con las categorías superiores.

1)

$n = 20$ datos

$x \equiv$ horas de estudio

$$y = \begin{cases} 1 & \text{aprueba examen} \\ 0 & \text{no aprueba examen} \end{cases}$$

X	Y
0.5	0
0.75	0
$\vdots$	$\vdots$
5	1
5.5	1

Modelo de regresión logística: $\hat{\theta}_0 = -4.072$ y $\hat{\theta}_1 = 1.5046$

a) ¿Como se interpreta $\hat{\theta}_1$?

$$\log \frac{p}{1-p} = \hat{\theta}_0 + \hat{\theta}_1 \cdot x$$
$$P \equiv Pr[Y = 1]$$
$$\log \frac{p}{1-p} = \exp(\hat{\theta}_0 + \hat{\theta}_1 x)$$
$$p = (1-p)\exp(\hat{\theta}_0 + \hat{\theta}_1 x)$$

b) $$Pr[Y = 1/X = 2] = \frac{1}{1 + \exp(-(-40.77 + 1.5046 \cdot 2))} \simeq 0.25$$
$$Pr[Y = 1/X = 3] = \frac{1}{1 + \exp(-(-40.77 + 1.5046 \cdot 3))} \simeq 0.607$$
$$\frac{Pr[Y = 1/X = 3]}{Pr[Y = 1/X = 2]} = \frac{0.607}{0.2556} = 2.428$$