

# Análisis Estadístico Multivariante

Francisco Javier Mercader Martínez

## Índice

<b>1</b>	<b>Vectores aleatorios</b>	<b>1</b>
1.1	Introducción . . . . .	1
1.2	Independencia de las variables aleatorias . . . . .	1
1.3	Distribuciones marginales . . . . .	1
1.4	Vector aleatorio absolutamente continuo . . . . .	2
1.5	Vector aleatorio discreto . . . . .	2
1.6	Distribuciones marginales . . . . .	2
1.6.1	Caso continuo . . . . .	2
1.6.2	Caso discreto . . . . .	3
1.7	Distribuciones condicionadas . . . . .	3
1.7.1	Caso continuo . . . . .	3
1.7.2	Caso discreto . . . . .	4
1.8	Distribución normal multivariante $\mathcal{N}_k(\mu, V)$ . . . . .	4
1.8.1	Normal bivalente . . . . .	5
1.9	Distribución multinomial $\mathcal{M}_k(n, p_1, \dots, p_k)$ . . . . .	9
1.10	Estadístico de Pearson . . . . .	9
1.11	Medias y covarianzas . . . . .	10
1.11.1	Esperanza de la transformación $g: \mathbb{R}^k \rightarrow \mathbb{R}$ . . . . .	10
1.12	Correlación . . . . .	11
1.12.1	Correlación entre vectores aleatorios . . . . .	12
1.13	Resultados básicos de la inferencia . . . . .	13
1.13.1	¿Cómo se representan las muestras aleatorias? . . . . .	13
1.13.2	¿Cómo se muestran los valores muestrales? . . . . .	13
1.13.3	El conjunto de datos <b>LifeCycleSavings</b> . . . . .	14
1.14	Estimador para el vector de medias $\mu$ . . . . .	14
1.14.1	¿Dónde se encuentra el vector de medias muestrales? . . . . .	15
1.15	Estimador para la matriz de covarianzas $V$ . . . . .	15
1.15.1	Para una distribución normal . . . . .	15
<b>2</b>	<b>Regresión lineal simple y múltiple</b>	<b>23</b>
2.1	Estimación de los parámetros . . . . .	23
2.1.1	Muestra . . . . .	23
2.2	Conjunto de datos <b>USArrests</b> . . . . .	23
2.2.1	Objetivo . . . . .	24
2.2.2	Matriz de gráficos . . . . .	25
2.2.3	Resumen numérico de las variables . . . . .	25

2.2.4	Estimación de las varianzas y covarianzas . . . . .	26
2.2.5	Modelo completo . . . . .	26
2.3	Regresión lineal simple . . . . .	26
2.3.1	Modelo teórico . . . . .	26
2.3.2	Función óptima . . . . .	27
2.4	Caso de normalidad . . . . .	27
2.4.1	Coeficiente de correlación de Pearson . . . . .	28
2.5	Caso de no normalidad . . . . .	28
2.6	Restricción sobre la función $h$ . . . . .	28
2.7	Minimizar la función costo . . . . .	29
2.7.1	Ecuaciones normales de la recta . . . . .	29
2.8	Expresión de la recta . . . . .	29
2.8.1	Recta de regresión para predecir $Y$ en función de $X$ . . . . .	30
2.9	Descomposición de la varianza . . . . .	30
2.9.1	Relaciones entre las varianzas . . . . .	30
2.10	Coeficiente de determinación . . . . .	30
2.11	Inferencia y predicción . . . . .	31
2.12	Función costo empírica . . . . .	31
2.12.1	Objetivo . . . . .	31
2.12.2	Diferenciar $J$ . . . . .	31
2.12.3	Solución exacta . . . . .	32
2.13	Regresión lineal múltiple . . . . .	33
2.13.1	Modelo teórico . . . . .	33
2.13.2	Obtención del mínimo . . . . .	33
2.14	Coeficiente de correlación múltiple . . . . .	33
2.14.1	Desigualdad de Cauchy-Schwarz . . . . .	34
2.14.2	Consecuencia . . . . .	34
2.15	Selección de variables . . . . .	35
2.15.1	Una opción . . . . .	35
2.15.2	Otra opción . . . . .	35
2.15.3	Otra opción más sencilla . . . . .	36
2.16	Inferencia y predicción . . . . .	36

# 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  $X = (X_1, \dots, X_k)$  tal que

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

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

### • Función de distribución conjunta

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

$$F(x_1, \dots, x_k) := P[X_1 \leq x_1, X_2 \leq x_2, \dots, X_k \leq x_k],$$

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

## 1.2) Independencia de las variables aleatorias

### • Definición

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

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

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

Esto es equivalente a que

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

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

## 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:

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

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

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

## 1.4) Vector aleatorio absolutamente continuo

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

$$F(x) = F(x_1, \dots, x_k) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_k} f(z_1, \dots, z_k) dz_k, \dots, dz_1,$$

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

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

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

Existen variables aleatorias cuya función de distribución es continua pero que no son absolutamente continuas (tienen una parte singular) y puede ocurrir que  $X_1, \dots, X_k$  sean absolutamente continuas y que  $(X_1, \dots, X_k)$  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 un vector aleatorio discreto:

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

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

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

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

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

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

con

$$f_{X_i} = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f(z_1, \dots, z_k) dz_1, \dots, dz_{i-1} \cdot dz_{i+1}, \dots, dz_k,$$

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

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

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

## 1.6.2) Caso discreto

- Distribución marginal de la variable aleatoria  $X_i$

Sea  $X = (X_1, \dots, X_l)$  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:

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

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

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

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

Nota:

$A$  y  $B$  independientes  $\longleftrightarrow P(A|B) = P(A)$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = P(A) \longrightarrow P(A \cap B) = P(A) \cdot P(B)$$

## 1.7) Distribuciones condicionadas

### 1.7.1) Caso continuo

- Distribución condicionada al valor de una variable

Sea  $X = (X_1, \dots, X_k)$  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  $(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_k)$  a  $(X_i = x_i^*)$  como la determinada por la función de densidad:

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

- Distribución condicionada a valores de varias variables

Sea  $X = (X_1, \dots, X_k)$  un vector aleatorio continuo con función de densidad  $f$ .

Sea  $(X_{i_1}, \dots, X_{i_m})$  un subvector arbitrario y  $(x_{i_1}^*, \dots, x_{i_m}^*) \in \mathbb{R}^m$  tal que:

$$f_{X_{i_1}, \dots, X_{i_m}}(x_{i_1}^*, \dots, x_{i_m}^*) > 0.$$

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

$$f_{X_1, \dots, X_{i_1-1}, X_{i_1+1}, \dots, X_{i_m-1}, \dots, X_k | X_{i_1} = x_{i_1}^*, \dots, X_{i_m} = x_{i_m}^*}(x_1, \dots, x_{i_1-1}, x_{i_1+1}, \dots, x_{i_m-1}, x_{i_m+1}, \dots, x_k | x_i^*) = \frac{f(x_1, \dots, x_{i_1}, \dots, x_{i_m}, \dots, x_k)}{f_{X_{i_1}, \dots, X_{i_m}}(x_{i_1}^*, \dots, x_{i_m}^*)}$$

## 1.7.2) Caso discreto

### • Distribución condicionada al valor de una variable

Sea  $X = (X_1, \dots, X_k)$  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  $(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_k)$  a  $(X_i = x_i^*)$  como la determinada por la función masa de probabilidad:

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

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

### • Distribución condicionada a valores de varias variables

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

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

$$P[X_{i_1} = x_{i_1}^*, \dots, X_{i_m} = x_{i_m}^*] > 0.$$

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

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

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

## 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  $\mu$  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?

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

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

### 1.8.1) Normal bivalente

- Función de densidad

Caso bivalente,  $k = 2$ , 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:

```
1 library("mvtnorm")
2 V <- matrix(c(1, 1/2,
3             1/2, 1), nrow = 2, ncol = 2, byrow = TRUE)
4 mu <- c(0, 0)
5 x <- c(1, 1)
6 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)**

```
1 library("mvtnorm")
2 V <- matrix(c(1, 1/2,
3             1/2, 1), nrow = 2, ncol = 2, byrow = TRUE)
4 mu <- c(0, 0)
5 x <- c(1, 1)
6 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

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

```

1 library("mvtnorm")
2 V <- matrix(c(1, 1/2,
3             1/2, 1), nrow = 2, ncol = 2, byrow = TRUE)
4 mu <- c(0, 0)
5 x1 <- c(-1, -1)
6 x2 <- c(1, 1)
7 pmvnorm(lower = x1, upper = x2, mean = mu, sigma = V)

```

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

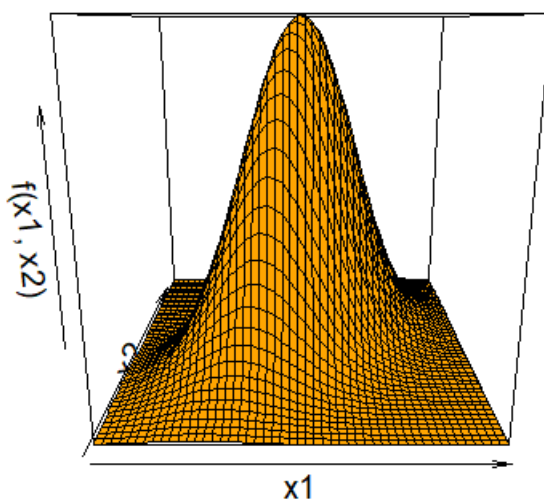
Su representación gráfica:

```

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

```

## Función de densidad



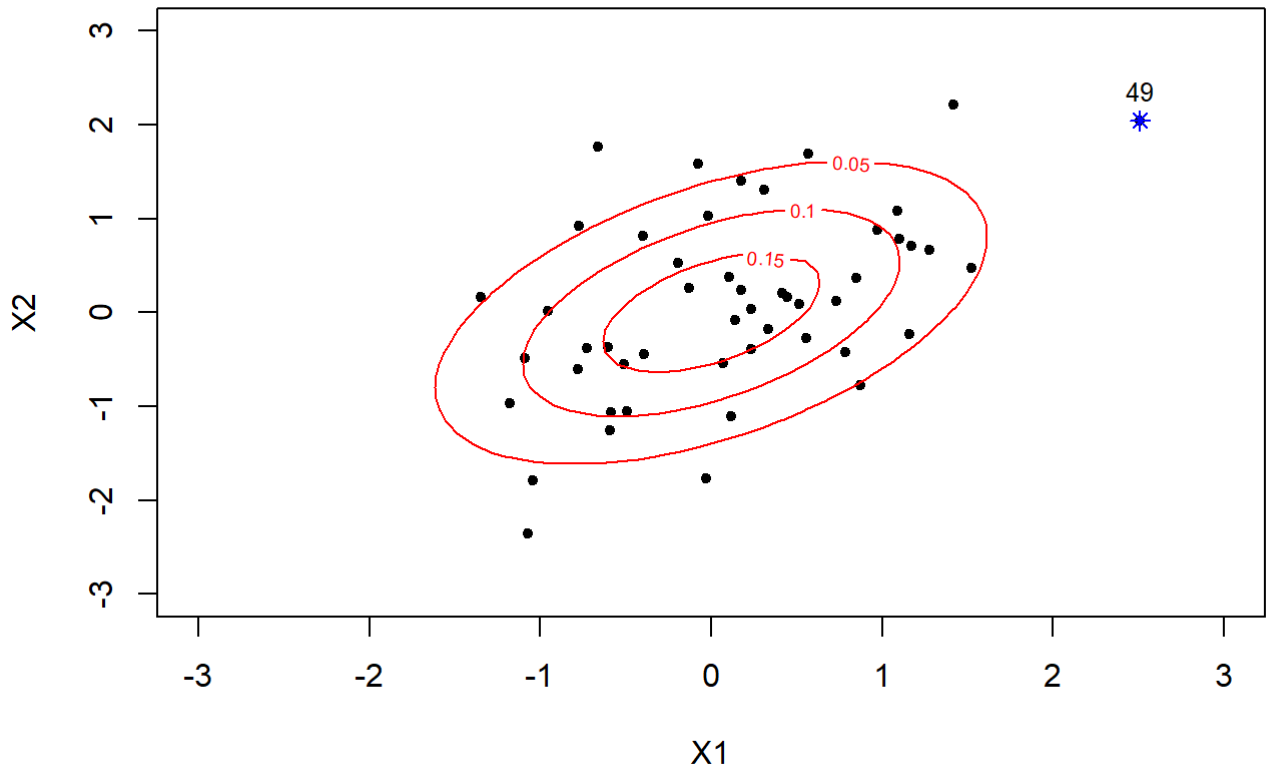
Su representación gráfica ( $f(x_1, x_2) = c$ ) y 50 datos simulados de este modelo

```

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

```





#### • Distancia de Mahalanobis

La distancia de Mahalanobis del vector  $x$  al vector  $\mu$  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](#).

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

#### • Distancias de los datos simulados al vector de medias teóricas $\mu$ con respecto a $V$

```
1 summary(dM1)
```

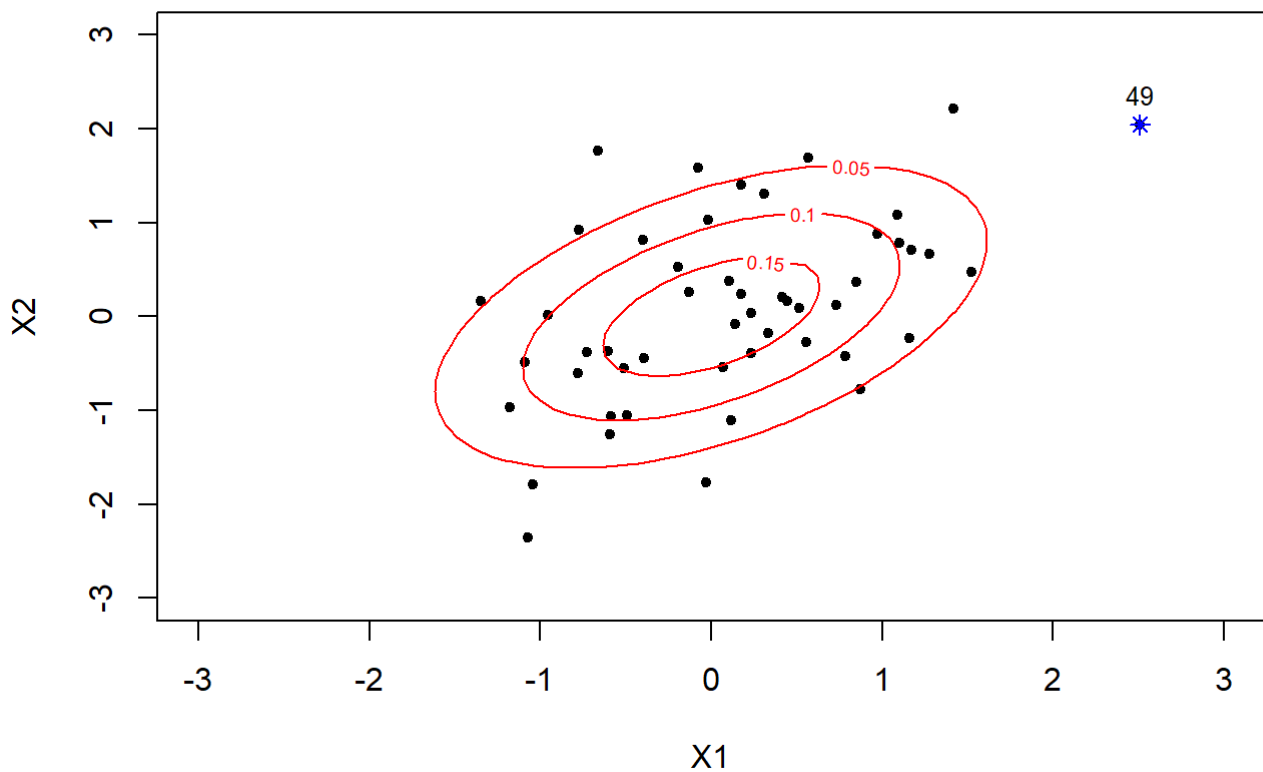
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.05216 0.41016 1.26433 1.66615 2.31591 7.13332
```

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

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

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

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



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

```
1 summary(dM2)
```

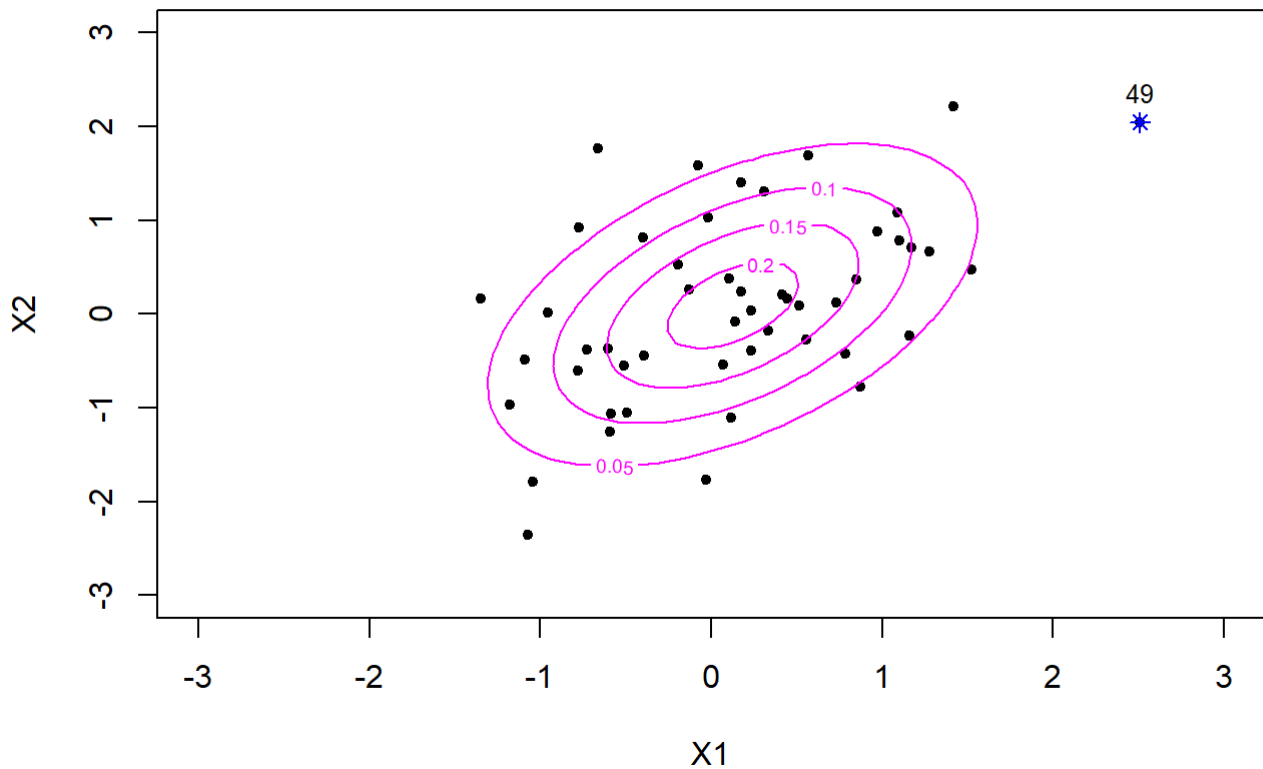
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.02114 0.67111 1.52636 1.96000 2.64131 8.65906
```

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

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

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

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



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

### • Modelo multinomial

$(X_1, \dots, X_k)$ : 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  $\{A_1, \dots, A_k\}$  y con probabilidades constantes  $p_i = P(A_i)$ , para  $i = 1, \dots, k$ .

**Función masa de probabilidad conjunta:**

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

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

**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,  $p_i = \frac{1}{6}$  para todo  $i$ , y los valores esperados son  $np_i = 10$ , para  $i = 1, \dots, 6$ .

**Objetivo:** Medir las discrepancias entre valores observados y esperados.

Sea  $X = (X_1, \dots, X_k)$  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  $\chi_{k-1}^2$  de Pearson con  $k - 1$  grados de libertad, cuando  $n \rightarrow \infty$ .

## 1.11) Medias y covarianzas

- Definiciones

Dado el vector aleatorio.

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

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

(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:

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

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

- 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:**

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

→ **Caso continuo:**

$$\begin{aligned} E[X_i] &= \int_{\mathbb{R}} x_i f_{X_i}(x_i) dx_i \\ &= \int_{\mathbb{R}^k} x_i f(x_1, \dots, x_k) dx_1 \cdots dx_k \end{aligned}$$

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

- Caso discreto

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

Si  $X$  es de tipo discreto,

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

Y en caso de existir:

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

- **Caso continuo**

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

Si  $X$  es de tipo continuo,

$$\exists E[g(X)] \longleftrightarrow \int_{\mathbb{R}^k} |g(x_1, \dots, x_k)| f_X(x_1, \dots, x_k) dx_1 \cdots dx_k < \infty$$

Y en caso de existir:

$$E[g(X_1, \dots, X_k)] = \int_{\mathbb{R}^k} g(x_1, \dots, x_k) f_X(x_1, \dots, x_k) dx_1 \cdots dx_k.$$

- **Propiedades**

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

En forma matricial,

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

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,  $\text{Cov}(X_i, X_j) = 0$ . El recíproco no es cierto.

Si  $X \rightarrow \mathcal{N}_k(\mu, V)$ , se puede demostrar que  $\mu$  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

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

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

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

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

y su [matriz de correlaciones](#) como

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

Puede demostrarse que

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

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

#### • Propiedades

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

- 1)  $E[a_1 g_1(X) + a_2 g_2(X)] = a_1 E[g_1(X)] + a_2 E[g_2(X)]$ , 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)  $\text{Cov}(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j]$ .
- 6) Si  $X_1, \dots, X_k$  son independientes,  $\text{Cov}(X_i, X_j) = 0$ .
- 7)  $\text{Var}(X_i + X_j) = \text{Var}(X_i) + 2\text{Cov}(X_i, X_j) + \text{Var}(X_j)$ .
- 8)  $\text{Cov}(aX_i + b, cX_j + d) = ac\text{Cov}(X_i, X_j)$ , donde  $a, b, c, d \in \mathbb{R}$ .
- 9)  $\text{Cov}(X) = E[(X - \mu)(X - \mu)'] = E[XX'] - \mu\mu'$ .
- 10)  $\text{Var}(a'X) = a'\text{Cov}(X)a = \sum_{i,j} a_i a_j \sigma_{i,j}$ , donde  $a \in \mathbb{R}^k$ .
- 11)  $\text{Cov}(AX + b) = A\text{Cov}(X)A'$ , donde  $A \in M_{m,k}$  y  $b' \in \mathbb{R}^m$ .
- 12) Si  $X_1, \dots, X_k$  son independientes,  $\text{Corr}(X_i, X_j) = 0$ .
- 13)  $\text{Corr}(aX_i + b, cX_j + d) = \text{Corr}(X_i, X_j)$ , donde  $a, b, c, d \in \mathbb{R}$ .
- 14)  $-1 \leq \text{Corr}(X_i, X_j) \leq 1$ .
- 15)  $\text{Corr}(X_i, aX_i + b) = \pm 1$ , donde  $a, b \in \mathbb{R}$  (según el signo de  $a$ ).
- 16)  $\text{Corr}(X) = \delta^{-1} \text{Cov}(X) \delta^{-1}$ , donde  $\delta$  es la matriz diagonal formada por las desviaciones típicas ( $\delta = \text{diag}(\sigma_1, \dots, \sigma_k)$ ).
- 17)  $\text{Cov}(X, Y) = (\text{Cov}(X_i, Y_j)) = \text{Cov}(Y, X)'$ .
- 18)  $\text{Cov}(X + Z, Y) = \text{Cov}(X, Y) + \text{Cov}(Z, Y)$

- 19) Si  $X$  e  $Y$  tienen la misma dimensión, entonces  $\text{Cov}(X + Y) = \text{Cov}(X) + \text{Cov}(X, Y) + \text{Cov}(Y, X) + \text{Cov}(Y)$ .
- 20)  $\text{Cov}(AX, BY) = A \text{Cov}(X, Y) B'$ , donde  $A$  y  $B$  son matrices (de dimensiones adecuadas).
- 21) Si  $X$  e  $Y$  independientes, entonces  $\text{Cov}(X, Y) = 0$ .

• Demostración apartado (10)

Directamente se tiene que:

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

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

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

### 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.

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

Si no hay grupos supondremos que los objetos

$$\mathbf{O}_i = (X_{i,1}, X_{i,2}, \dots, X_{i,k})'$$

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

#### 1.13.2) ¿Cómo se muestran los valores muestrales?

• Matriz de datos

En general, nuestra muestra se representará como:

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

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

```
1 datos <- LifeCycleSavings
2 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 $\mu$

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

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

donde  $\bar{X}_j = \frac{1}{n} \sum_{i=1}^n X_{i,j}$ .

Se puede demostrar fácilmente que:

$$\begin{aligned} \rightarrow E(\bar{O}) &= \mu \text{ (estimador centrado de } \mu) \\ \rightarrow \text{Cov}(\bar{O}) &= \frac{V}{n} \end{aligned}$$

Si no hay grupos, supondremos que los vectores

$$O_i = (x_{i,1}, x_{i,2}, \dots, x_{i,k})'$$

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.



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

- Propiedad

$\bar{O}$  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

$$\min_{P \in \mathbb{R}^k} MSE = \sum_{i=1}^n d^2(O_i, P),$$

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

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

$\rightarrow$  La cuasi-covarianza muestral:

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

Para estimar  $V$  usaremos:

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

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

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

#### 1.15.1) Para una distribución normal

- Proposición

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

- $\bar{O} \rightarrow \mathcal{N}_k\left(\mu, \frac{V}{n}\right)$
- $\bar{O}$  y  $\hat{V}$  son los **estimadores máximos verosímiles** de  $\mu$  y  $V$ , respectivamente.
- Además,  $\bar{O}$  y  $\hat{V}$  son **independientes entre sí**. Por tanto, también  $\bar{O}$  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.

$$H_0 : (X_1, \dots, X_k) \rightarrow \mathcal{N}_k(\mu, V)$$

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

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 bivalente lógicamente obtendremos un  $p$ -valor que apoya la hipótesis nula.

```
1 library("mvnormtest")
2 V <- matrix(c(1, 1/2,
3               1/2, 1), nrow = 2, ncol = 2, byrow = TRUE)
4 mu <- c(0, 0)
5 seed = set.seed(2023)
6 d <- rmvnorm(50, mu, V)
7 mshapiro.test(t(d))
```

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

#### • Seguimos con `LifeCycleSavings`

Cálculo de las medias muestrales para cada variable.

```
1 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

```
1 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

```
1 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

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy = \int_{-\infty}^0 0 dy + \int_0^1 1 dy + \int_1^{+\infty} 0 dy = [y]_{y=0}^{y=1} = 1 \quad f_X(x) = \begin{cases} 1 & \text{si } 0 < x < 1 \\ 0 & \text{en otro caso} \end{cases}$$

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

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

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

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

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

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

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

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

Los recintos son dependientes.

$$y|x = x^*$$

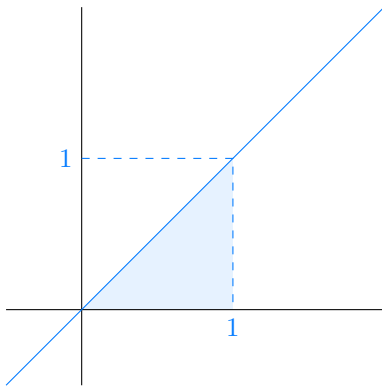
$$f_X(x^*) > 0$$

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

$$x|y = y^*$$

$$f_Y(y^*) > 0$$

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



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  $X = \frac{1}{2}$ .

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

$$y|x = x^*$$

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

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

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.

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

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

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

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:

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

$$\begin{aligned}
\mu_x = \mu_y &= 0 \\
\sigma_x^2 &= 1 \\
\sigma_y^2 &= 4 \\
\rho &= \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{1}{\sqrt{1} \cdot \sqrt{4}} = \frac{1}{2}
\end{aligned}
\quad
\begin{aligned}
f(1,1) &= \frac{1}{2\pi \cdot 1 \cdot 2\sqrt{1 - (\frac{1}{2})^2}} \exp\left(-\frac{1}{2\left(1 - (\frac{1}{2})^2\right)} \cdot \left[1^2 + \frac{1^2}{4} - \frac{2 \cdot \frac{1}{2}}{1 \cdot 2}\right]\right) \\
&= \frac{1}{2\pi\sqrt{3}} \exp\left(-\frac{2}{3} \cdot \frac{3}{4}\right) \\
&= \frac{1}{2\pi\sqrt{3}} \exp\left(-\frac{1}{2}\right) \simeq \boxed{0.0557}
\end{aligned}$$

$$f(x) = \frac{1}{|V|(2\pi)^k} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}$$

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

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

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

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

6) Sea  $(X, Y)$  un vector aleatorio con distribución uniforme en el 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,  $E[XY^2]$ .

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

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

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:

$$E[XY^2] = \int_0^1 \int_0^1 xy^2 \, dx \, dy$$

Resolviendo la integral obtenemos:

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

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

$$E[XY^2] = E[X] \cdot E[Y^2] = (*) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$$

$$E[X] = \int_0^1 x \, dx = \left[ \frac{x^2}{2} \right]_0^1 = \frac{1}{2}$$

$$E[Y^2] = \int_0^1 y^2 \, dy = \left[ \frac{y^3}{3} \right]_0^1 = \frac{1}{3}$$

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

X \ Y	1	2
	1	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:

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

$$\text{Var}(X) = E[X^2] - (E[X])^2 = 3 - \left(\frac{5}{3}\right)^2 = \frac{2}{9}$$

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y] = \frac{25}{9} - \frac{5}{3} \cdot \frac{5}{3} = 0$$

$$\text{Var}(Y) = E[Y^2] - (E[Y])^2 = 3 - \left(\frac{5}{3}\right)^2 = \frac{2}{9}$$

$$E[X^2] = 1^2 \cdot P(X = 1) + 2^2 \cdot P(X = 2) = 1^2 \cdot \left(\frac{1}{9} + \frac{2}{9}\right) + 2^2 \cdot \left(\frac{2}{9} + \frac{4}{9}\right) = \frac{1}{3} + \frac{8}{3} = 3$$

$$E[Y^2] = 1^2 \cdot P(Y = 1) + 2^2 \cdot P(Y = 2) = 1^2 \cdot \left(\frac{1}{9} + \frac{2}{9}\right) + 2^2 \cdot \left(\frac{2}{9} + \frac{4}{9}\right) = \frac{1}{3} + \frac{8}{3} = 3$$

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

- Matriz de correlaciones:

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

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

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  $X_1, \dots, X_k$ .

Para ello usaremos una función predictora lineal

$$h_{\theta}(x_1, \dots, x_k) := \theta_0 + \theta_1 x_1 + \dots + \theta_k x_k,$$

donde  $\theta = (\theta_0, \dots, \theta_k)' \in \mathbb{R}^{k+1}$  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

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

que equivale a la notación introducida en el tema anterior

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

Los datos se representan en forma de tabla:

$i$	$x_1$	$x_2$	$\dots$	$x_k$	$y$	$i$	$x_1$	$x_2$	$\dots$	$x_k$	$y$
$\mathbf{O_1}$	$x_1^{(1)}$	$x_2^{(1)}$	$\dots$	$x_k^{(1)}$	$y_1$	$\mathbf{O_1}$	$x_{1,1}$	$x_{1,2}$	$\dots$	$x_{1,k}$	$y_1$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\mathbf{O_i}$	$x_1^{(i)}$	$x_2^{(i)}$	$\dots$	$x_k^{(i)}$	$y_i$	$\mathbf{O_i}$	$x_{i,1}$	$x_{i,2}$	$\dots$	$x_{i,k}$	$y_i$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\mathbf{O_n}$	$x_1^{(n)}$	$x_2^{(n)}$	$\dots$	$x_k^{(n)}$	$y_n$	$\mathbf{O_n}$	$x_{n,1}$	$x_{n,2}$	$\dots$	$x_{n,k}$	$y_n$

## 2.2) Conjunto de datos USArrests

### Cargamos los datos

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

```

1 d <- USArrests
2 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 ( $Y = \text{Murder}$ ) en función de la variable  $X = \text{UrbanPop}$ .

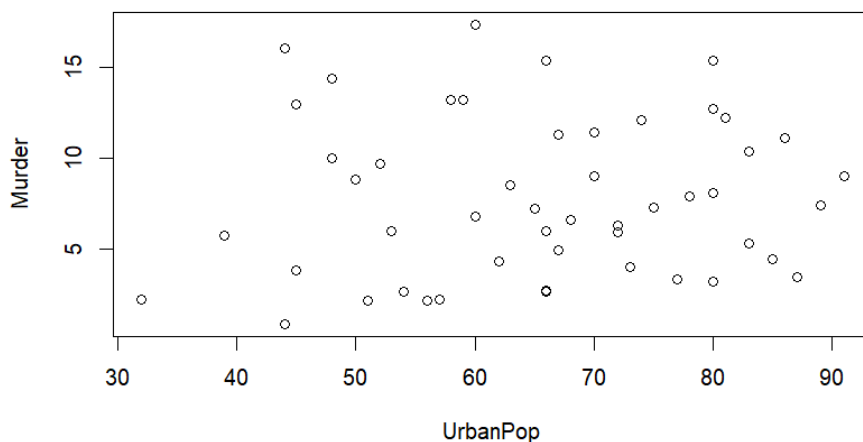
Para visualizar la relación entre estas variables podemos representarlas situando  $X$  en el eje horizontal e  $Y$  en el vertical.

```

1 x <- d$UrbanPop #Elegimos x
2 y <- d$Murder   #Elegimos y
3 plot(x, y, xlab = 'UrbanPop', ylab = 'Murder')

```

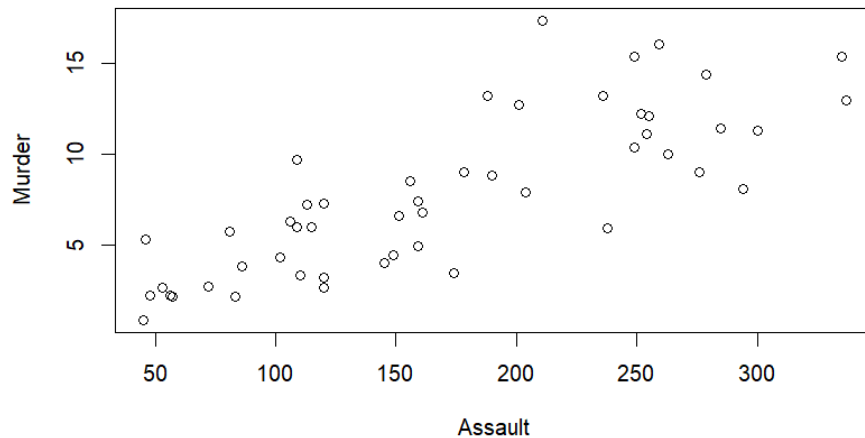
Podemos observar que no parece existir ninguna relación entre las variables **Murder** y **UrbanPop** por lo que la predicción no será muy buena.



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

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

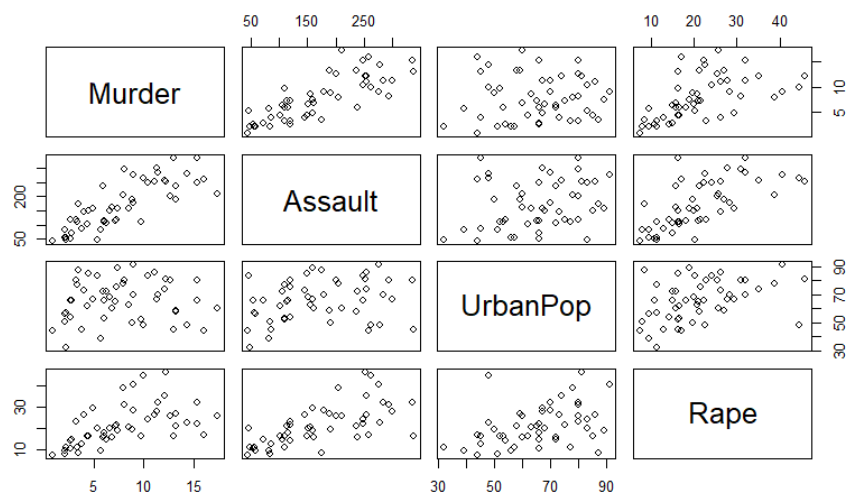
```
1 x <- d$Assault #Elegimos x
2 y <- d$Murder #Elegimos y
3 plot(x, y, xlab = 'Assault', ylab = 'Murder')
```



### 2.2.2) Matriz de gráficos

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

```
1 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.

```
1 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).

```

1 #calculo directo de la varianza y cuasivarianza para Murder
2 mu <- mean(d$Murder)
3 n <- length(d)
4 hat_sigma = sum((d$Murder-mu)^ 2)/n #varianza muestral
5 S = sum((d$Murder-mu)^ 2)/(n-1) #cuasivarianza muestral
6 #calculo de la matriz de covarianzas
7 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:

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

(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

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

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

$$\begin{aligned}\mu_1 &= \mu_X = E[X] \\ \mu_2 &= \mu_Y = E[Y]\end{aligned}$$

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

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

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

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

con

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

- Observaciones

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

Los residuos condicionados  $R_x = Y|X=x$  también serán normales  $R_x \rightarrow \mathcal{N}(0, \bar{\sigma}^2)$  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  $\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$  es el **coeficiente de correlación lineal de Pearson**.

La recta siempre pasa por el punto  $(\mu_X, \mu_Y)$ .

### 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:

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

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,

$$h_{\theta}(x) := \theta_0 + \theta_1 x$$

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

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

donde

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

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,

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

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

Con lo cual,

$$\rightarrow \hat{\theta}_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{\sigma_{X,Y}}{\sigma_X^2}$$

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

## 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  $\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$  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:

$$\rightarrow E(\hat{Y}) = E(Y)$$

$$\rightarrow E(R) = 0$$

## 2.9) Descomposición de la varianza

### 2.9.1) Relaciones entre las varianzas

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

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

Puesto que

$$\begin{aligned} \text{Var}(\hat{Y}) &= \hat{\theta}_1^2 \sigma_X^2 = \frac{\sigma_{X,Y}^2}{\sigma_X^2} = \rho_{X,Y}^2 \sigma_Y^2 \\ \text{Var}(R) &= \text{Var}(Y - \hat{\theta}_1 X) = (1 - \rho_{X,Y}^2) \sigma_Y^2. \end{aligned}$$

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

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

donde  $n$  será el tamaño muestral.

Los datos de cada variable se representarán como columnas y todos los 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,

$$h_{\theta} := \theta_0 + \theta_1 x,$$

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

Para calcular estos parámetros minimizaremos una función costo 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

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

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

### 2.12.2) Diferenciar $J$

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

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

Igualando a cero obtenemos las [ecuaciones normales empíricas](#):

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

### 2.12.3) Solución exacta

De la primera ecuación,

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

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

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

De la segunda,

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

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

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

Obtenemos:

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

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

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

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

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

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  $X_1, \dots, X_k$  (sobre el mismo espacio de probabilidad).

En el modelo lineal queremos construir una función

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

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

Para medir el error usaremos de nuevo el error cuadrático medio

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

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

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

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

### 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 mínimo  $\hat{\theta}' \in \mathbb{R}^{k+1}$ .

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

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

para  $j = 0, \dots, k$ , 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)

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

## 2.14) Coeficiente de correlación múltiple

- Definición

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

$$\text{Corr}^2(\mathbf{X}, Y) = \rho_{k+1, (1, \dots, k)}^2 = \frac{v'_{1,2} V_{\mathbf{X}}^{-1} v_{1,2}}{\sigma_Y^2}$$

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

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,

$$\rho_{2,(1)}^2 = \frac{\sigma_{1,2} \sigma_{1,1}^{-1} \sigma_{1,2}}{\sigma_{2,2}} = \rho_{1,2}^2$$

### • Proposición

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

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

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

$$\begin{aligned} \text{Corr}^2(Y, \alpha' \mathbf{X}) &= \frac{(\text{Cov}(Y, \alpha' \mathbf{X}))^2}{\sigma_Y^2 \text{Var}(\alpha' \mathbf{X})} = \frac{(\text{Cov}(Y, \mathbf{X}) \alpha)^2}{\sigma_Y^2 \text{Cov}(\alpha' \mathbf{X}, \alpha' \mathbf{X})} \\ &= \frac{(\alpha' v_{1,2})^2}{\sigma_Y^2 \alpha' V_{\mathbf{X}} \alpha} = \frac{(\alpha' V_{\mathbf{X}}^{\frac{1}{2}} V_{\mathbf{X}}^{-\frac{1}{2}} v_{1,2})^2}{\sigma_Y^2 \alpha' V_{\mathbf{X}} \alpha} \end{aligned}$$

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

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

es decir,  $\rho_{k+1, (1, \dots, k)}^2$  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  $\mathbf{X} = (X_1, \dots, X_k)'$  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  $i \neq j$ ,  $i, j \in \{1, \dots, k\}$ ,  $V_{\mathbf{X}}$  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  $\text{Cov}(Y, \mathbf{X}) = (\text{Cov}(Y, X_1), \dots, \text{Cov}(Y, X_k))$ .

## 2.15) Selección de variables

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

### 2.15.1) Una opción

Seleccionar la variable  $Z_1 = X_{j_1}$  que maximice la correlación al cuadrado con  $Y$ :

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

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

Seleccionar la variable  $Z_2 = X_{j_2}$  con  $j \neq j_1$  que más información tenga sobre ese residuo:

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

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 correlación 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 variables 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):  $(x_1^{(i)}, \dots, x_k^{(i)}, y^{(i)})$ , para  $i = 1, \dots, n$ , donde conocemos los valores de  $y$  para esos valores de  $x$ .
  - Los datos se colocarán como  $k + 1$  columnas (variables) y  $n$  filas

## RELACIÓN DE PROBLEMAS: REGRESIÓN LINEAL SIMPLE Y MÚLTIPLE

### 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

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

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  $\sum_i^n (y_i - \hat{y}_i)^2 = 209.55$ ,  $\sum_i^n (y_i - \bar{y})^2 = 6777.5$ .

Se pide:

- 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.
  - ¿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?
  - 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

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

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 Fahrenheit?

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  $f_X$ :

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

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

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

b) ¿Coincide con la recta de regresión?

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

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

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

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

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

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

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

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

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 los valores de  $X$  y proporcionar una predicción para  $X = \frac{2}{3}$

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

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

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

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

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

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

$$\mu = (1, 2)$$

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

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

- 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:  $y = \hat{\theta}_0 + \hat{\theta}_1 x$

$$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x} = 3.2 - (-0.36) \cdot 3 = 4.29$$

$$\hat{\theta}_1 = \frac{S_{xy}}{S_x^2} = -\frac{1.6}{4.4} = -0.36$$

$$n = 5$$

$$\bar{x} = 3$$

$$\bar{y} = 3.2$$

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

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

Recta de regresión:  $y = 4.29 - 0.36x \xrightarrow{x=3} y = 4.29 - 0.36 \cdot 3 = 3.2$

Coefficiente de correlación lineal al cuadrado:

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

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

- 5) En 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

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

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

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

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

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

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

- 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)  $x_3 \longrightarrow x_3 + 1$

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

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

Ajuste bueno.

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

$$\hat{\sigma}^2 = \frac{SCR}{n} = \frac{1950.7292}{6} = 325.0705$$

$$\tilde{\sigma}^2 = \frac{SCR}{n-k-1} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-k-1} = 650.141$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{1950.423}{14112} = 1 - 0.1382 = 0.8618$$

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