Análisis Estadístico Multivariante

Francisco Javier Mercader Martínez

Índice

1 Vectores aleatorios **1**

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 bivariante . . . 5
1.9 Distribución multinomial $\mathcal{M}_k(n, p_1, \ldots, p_k)$ . . . 9
1.10 Estadístico de Pearson . . . 9
1.11 Medias y covarianzas . . . 10
1.11.1 Esperanza de la transformación $g : \mathbb{R}^k \longrightarrow \mathbb{R}$ . . . 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 `LifeCycleSavings` . . . 14
1.14 Estimador para el vector de medias μ . . . 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

2 Regresión lineal simple y múltiple **24**

2.1 Estimación de los parámetros . . . 24
2.1.1 Muestra . . . 24
2.2 Conjunto de datos `USArrests` . . . 24
2.2.1 Objetivo . . . 25
2.2.2 Matriz de gráficos . . . 26
2.2.3 Resumen numérico de las variables . . . 26

2.2.4 Estimación de las varianzas y covarianzas . . . 27
2.2.5 Modelo completo . . . 27
2.3 Regresión lineal simple . . . 27
2.3.1 Modelo teórico . . . 27
2.3.2 Función óptima . . . 28
2.4 Caso de normalidad . . . 28
2.4.1 Coeficiente de correlación de Pearson . . . 29
2.5 Caso de no normalidad . . . 29
2.6 Restricción sobre la función h . . . 29
2.7 Minimizar la función costo . . . 30
2.7.1 Ecuaciones normales de la recta . . . 30
2.8 Expresión de la recta . . . 30
2.8.1 Recta de regresión para predecir Y en función de X . . . 30
2.9 Descomposición de la varianza . . . 31
2.9.1 Relaciones entre las varianzas . . . 31
2.10 Coeficiente de determinación . . . 31
2.11 Inferencia y predicción . . . 31
2.12 Función costo empírica . . . 32
2.12.1 Objetivo . . . 32
2.12.2 Diferenciar J . . . 32
2.12.3 Solución exacta . . . 32
2.13 Regresión lineal múltiple . . . 33
2.13.1 Modelo teórico . . . 33
2.13.2 Obtención del mínimo . . . 34
2.14 Coeficiente de correlación múltiple . . . 34
2.14.1 Desigualdad de Cauchy-Schwarz . . . 35
2.14.2 Consecuencia . . . 35
2.15 Selección de variables . . . 36
2.15.1 Una opción . . . 36
2.15.2 Otra opción . . . 36
2.15.3 Otra opción más sencilla . . . 36
2.16 Inferencia y predicción . . . 37
2.17 Función costo empírica . . . 37
2.17.1 Descomposición de la variabilidad . . . 38
2.17.2 Coeficiente de determinación: R^2 . . . 39
2.17.3 Propiedades . . . 39
2.17.4 Coeficiente de determinación ajustado . . . 39
2.18 Extensiones del modelo de regresión múltiple . . . 39
2.18.1 Planteamiento . . . 39
2.18.2 Problema de sobreajuste (`overfitting`) . . . 40

3 Regresión logística y multinomial **54**
3.1 Modelo de regresión logística . . . 54
3.1.1 Contexto . . . 54
3.1.2 Objetivo . . . 54
3.1.3 ¿Cómo elegir la función g? . . . 54
3.2 Función logística . . . 55
3.3 ¿Cómo determinar una función costo que penalice las decisiones erróneas? . . . 55

3.3.1 Función costo . . . 55
3.3.2 Criterio . . . 56
3.3.3 Otra formulación del problema . . . 56
3.4 Inferencia y predicción . . . 56
3.4.1 Función costo empírica . . . 56
3.4.2 Función costo empírica en forma matricial . . . 57
3.4.3 Objetivo . . . 57
3.5 Un ejemplo sencillo . . . 57
3.5.1 Datos muestrales . . . 57
3.6 Regresión logística multinomial . . . 62
3.6.1 Contexto . . . 62
3.6.2 Objetivo . . . 62
3.7 Modelo teórico . . . 62
3.7.1 Formulación . . . 62
3.7.2 Observaciones . . . 63
3.7.2.1) Un ejemplo ficticio . . . 63
3.7.2.2) Interpretación de los parámetros . . . 64
3.7.2.3) Estimador de máxima verosimilitud . . . 64
3.8 Criterio de máxima de verosimilitud para nuestro modelo . . . 64
3.8.1 Criterio . . . 64
3.8.2 Función de verosimilitud . . . 65
3.8.3 Función de log-verosimilitud . . . 65
3.8.4 ¿Cómo obtenemos en la práctica las estimaciones de $\theta_1, \ldots, \theta_{g-1}$? . . . 66
3.9 Un caso sencillo . . . 66
3.9.1 Cálculo de la verosimilitud . . . 66
3.9.2 Estimación de los parámetros . . . 70
3.9.3 Modelo estimado . . . 71
3.10 Modelo logístico multinomial con categorías ordinales . . . 71
3.10.1 Modelo teórico . . . 71

4 Análisis de componentes principales 76
4.1 Introducción . . . 76
4.2 ¿Cómo realizamos estos cálculo en `R`? . . . 84
4.3 Desigualdades . . . 85
4.3.1 Desigualdad de Chebyshev . . . 86
4.3.2 Desigualdad de Chebyshev multivariante . . . 86
4.4 Propiedades . . . 87
4.5 Cálculo a partir de la matriz de correlaciones . . . 89
4.5.1 Estimación de la matriz de covarianzas . . . 91
4.5.2 Cálculo a partir de una muestra . . . 91
4.5.3 Definiciones . . . 91
4.5.4 Cálculo a partir de la matriz de correlaciones . . . 91
4.5.5 Caso de muestras grandes . . . 92
4.5.6 Consecuencia . . . 92
4.5.7 Caso de normalidad . . . 92
4.5.8 Cálculo de las componentes principales maximizando la varianza muestral . . . 92
4.5.9 Interpretaicón geométrica: cálculo minimizando las distancias cuadráticas . . . 93

5 Análisis discriminante **101**
5.1 Introducción . . . 101
5.1.1 Objetivo . . . 101
5.1.2 Criterios . . . 101
5.1.3 Distancia de Mahalanobis . . . 101
5.1.4 Para una **normal bivariante** . . . 102
5.2 Dos poblaciones normales con la misma matriz de covarianzas . . . 102
5.2.1 Clasificación teórica . . . 102
5.2.2 Función discriminante de Fisher . . . 104
5.2.3 Regla de discriminación . . . 104
5.2.4 ¿Cómo de **buena** es la función discriminante de Fisher obtenida? . . . 105
5.2.5 Otro criterio para medir la bondad de un criterio de clasificación . . . 105
5.2.6 Diferente importancia a los dos tipos de errores . . . 108
5.2.7 Criterio de mínimo coste (probabilidad de error) . . . 109
5.2.8 Criterio de máxima probabilidad a posteriori . . . 109
5.3 Varias poblaciones con la misma matriz de covarianza . . . 109
5.4 Varias poblaciones con distintas matrices de covarianza . . . 112
5.4.1 Criterios de clasificación . . . 112
5.4.2 Observaciones . . . 112
5.5 Clasificación a partir de una muestra . . . 113

6 Análisis Cluster **124**
6.1 Introducción . . . 124
6.1.1 Objetivo . . . 124
6.1.2 Contexto . . . 124
6.2 Distancias entre individuos . . . 124
6.2.1 La distancia Euclídea . . . 124
6.2.2 La distancia de Mahalanobis . . . 125
6.2.3 Otras distancias interesantes . . . 126
6.3 Distancia de individuos a grupos y distancias entre grupos . . . 126
6.3.1 Distancias de individuos a grupos . . . 126
6.3.2 Función costo . . . 127
6.4 Métodos cluster . . . 127
6.4.1 Clasificación . . . 127
6.5 Método no jerárquico de las K-medias . . . 127
6.5.1 Algoritmo del método de las K-medias . . . 128
6.5.2 Un ejemplo sencillo . . . 129
6.5.3 ¿Cómo comparar distintas soluciones? . . . 136
6.5.4 `K-means` se puede ejecutar de forma automática en `R` . . . 138
6.6 Método jerárquico . . . 141
6.6.1 Índice de similaridad . . . 141
6.6.2 Algoritmos . . . 141
6.6.3 Un ejemplo sencillo . . . 141
6.6.4 ¿Cómo realizar este agrupamiento de forma automática en `R`? . . . 143

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

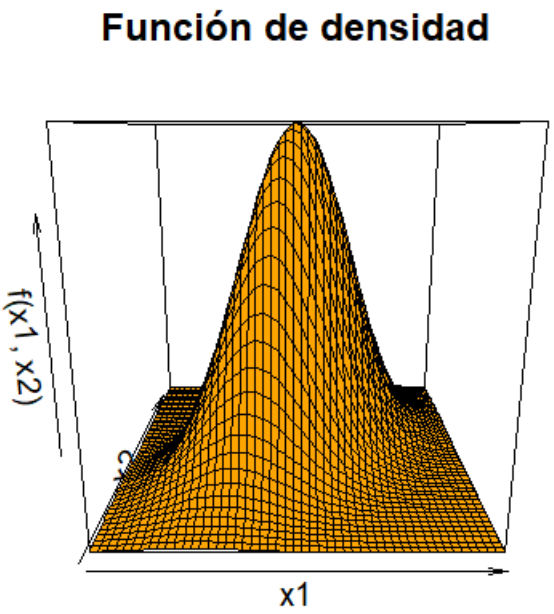
$$P(-1 < X_1 < 1,\ -1 < X_2 < 1)$$

```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
```

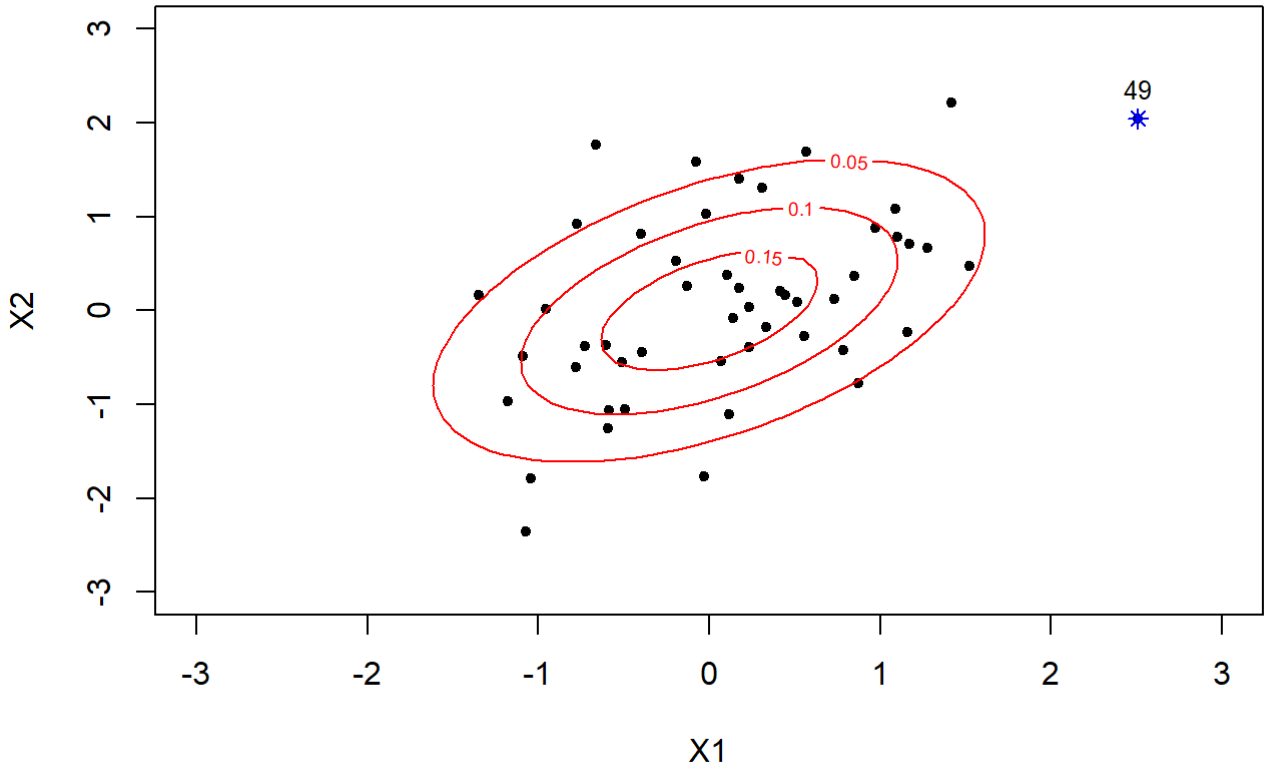
Su representación gráfica:

```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)
```

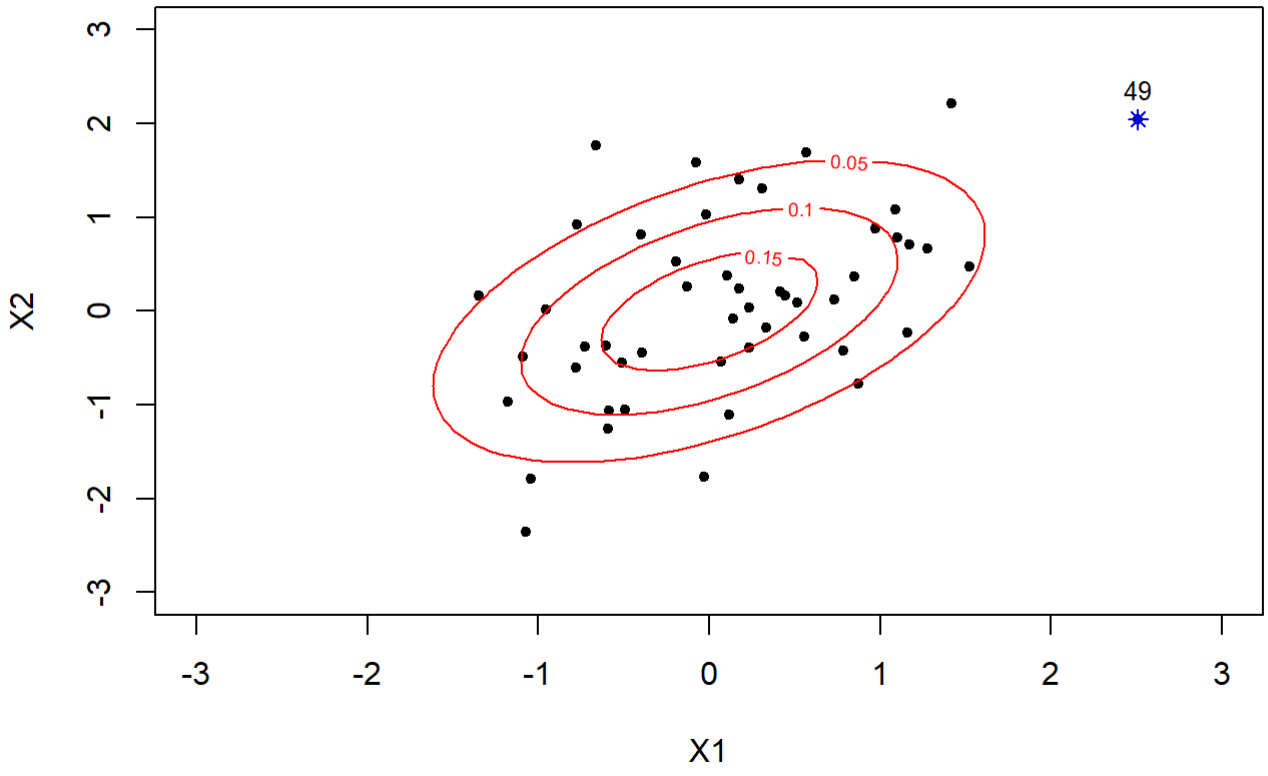
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.05216 0.41016 1.26433 1.66615 2.31591 7.13332
```

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

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

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

```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)
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 $\mathcal{S}$

```r
summary(dM2)
```

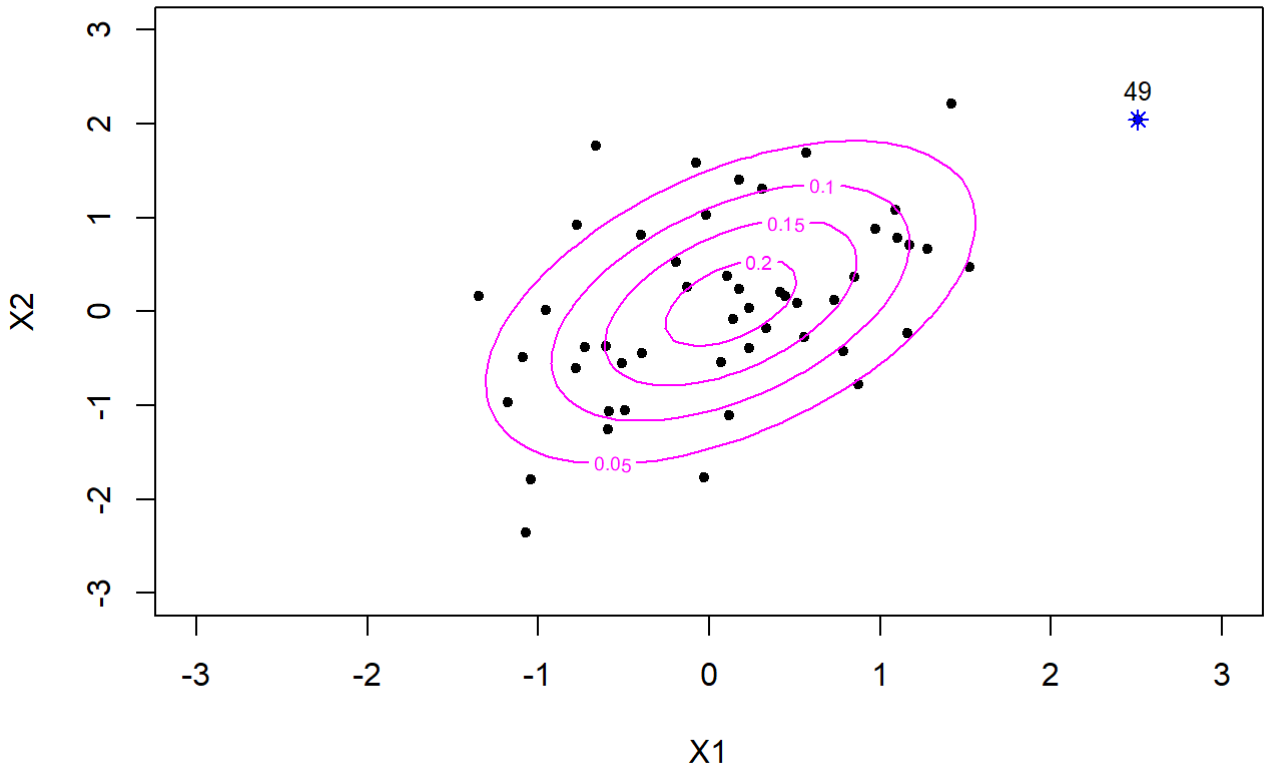
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.02114 0.67111 1.52636 1.96000 2.64131 8.65906
```

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

```r
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)

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

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

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

$\rightarrow$ 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:

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

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

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

```r
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

```r
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.

$y|x=x^*$

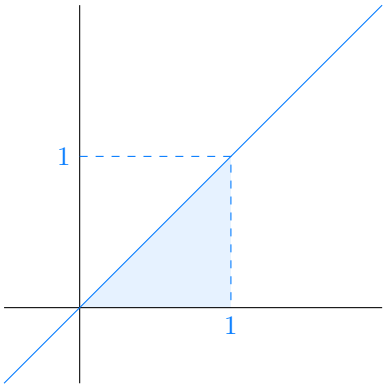
$f_X(x^*)>0$

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

$x|y=y^*$

$f_Y(y^*)>0$

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



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

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

$y|x = x^*$

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

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

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.

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

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

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

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:

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

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

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

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

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

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

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

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

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).

Dado un conjunto de puntos en $\mathbb{R}^k$, $\{x_1, x_2, \ldots, x_n\}$, la media muestral $\overline{x}$ se define como:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Queremos demostrar que $\overline{x}$ minimiza la suma de las distancias al cuadrado a todos los puntos en el conjunto, es decir, minimiza la función:

$$f(y) = \sum_{i=1}^{n} \|x_i - y\|^2$$

Para encontrar el mínimo de esta función, tomamos la derivada con respecto a y y la igualamos a cero:

$$\frac{\mathrm{d}f}{\mathrm{d}y} = \frac{\mathrm{d}}{\mathrm{d}y} \sum_{i=1}^{n} \|x_i - y\|^2 = -2 \sum_{i=1}^{n} (x_i - y) = 0$$

Resolviendo para y obtenemos:

$$y = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}$$

Por lo tanto, hemos demostrado que el vector de medias muestral $\overline{x}$ es el punto en $\mathbb{R}^k$ que minimiza la suma de las distancias al cuadrado a todos los puntos en el conjunto, es decir, minimiza el 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)`.

```
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.

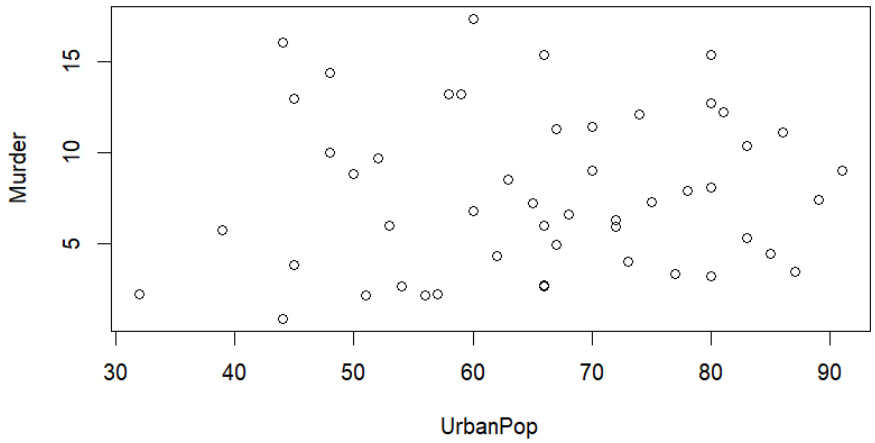
2.2.1) Objetivo

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.

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

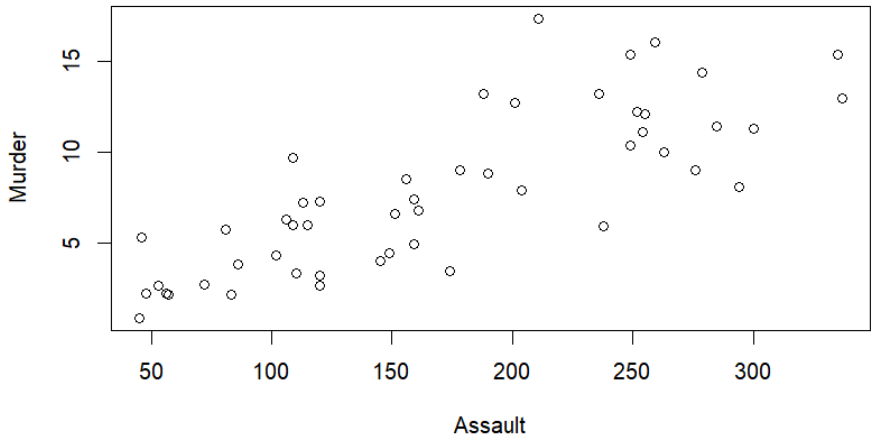
Podemos observar que no parece existir ninguna relación entre las variables `Murder` y `UrbanPop` por lo que la predicción no será muy buena.



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

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

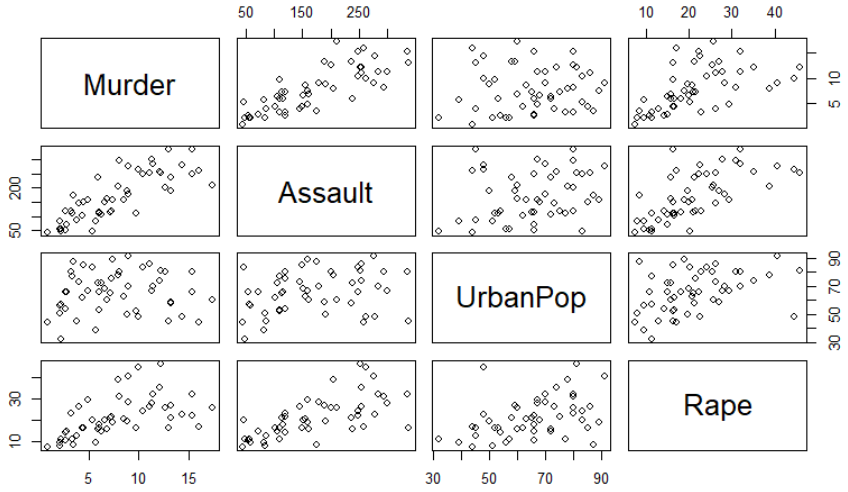
```
x <- d$Assault #Elegimos x
y <- d$Murder #Elegimos y
plot(x, y, xlab = 'Assault', ylab = 'Murder')
```



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

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

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

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

Obtenemos:

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

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

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

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

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

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 $\mathbf{X} = (X_0, X_1, \dots, X_k)'$ 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 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

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

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

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

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

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,

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

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

- Demostración

De la definición se tiene

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

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

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

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

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

es decir, si $\alpha = \lambda V^{-1}_{\mathbf{X}} v_{1,2}$ 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

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

- 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.

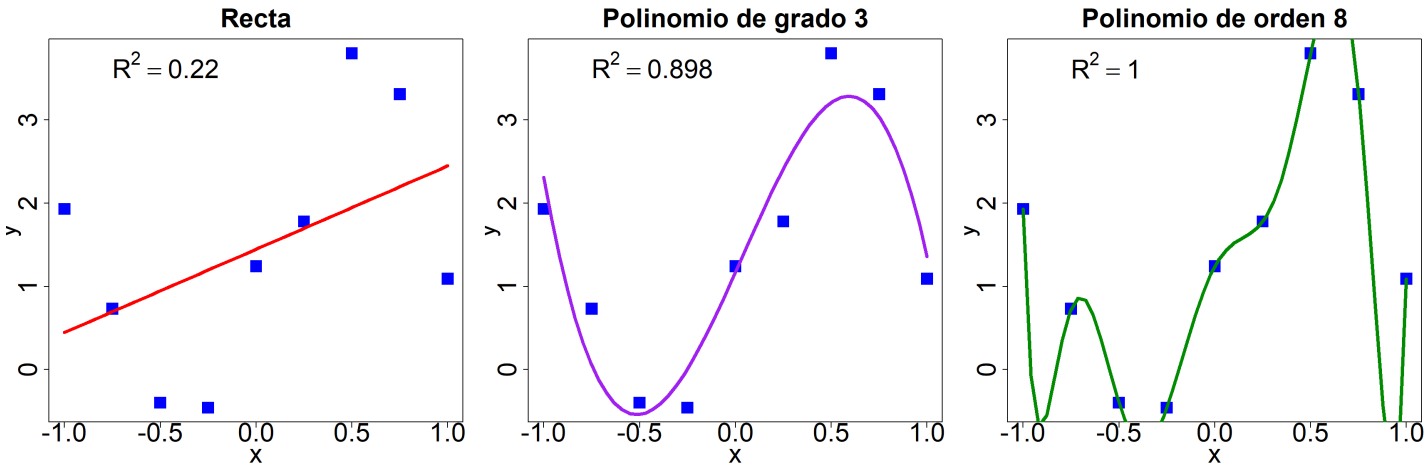
También se nos puede dar el caso contrario, denominado subajuste (`underfitting`)

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)
```

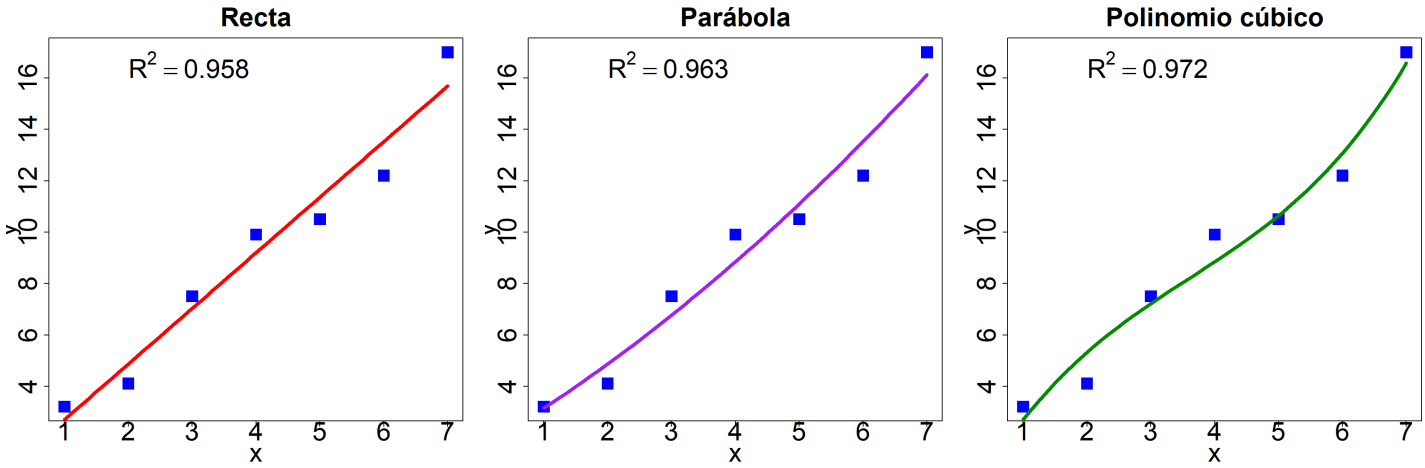


Otro ejemplo

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)
```

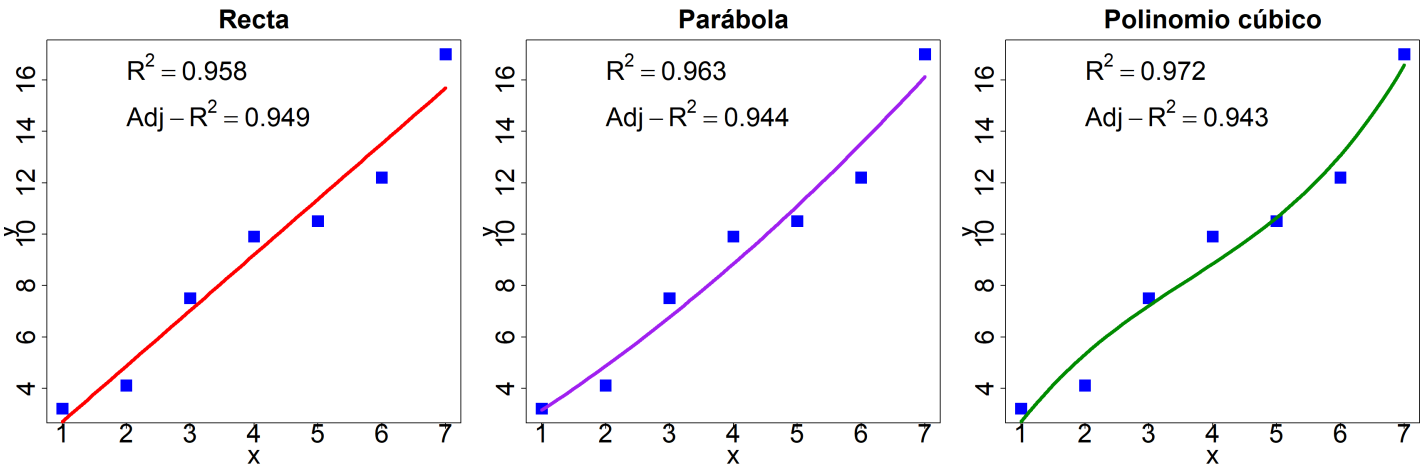


Para nuestro ejemplo

¿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)

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 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

$$y = 56.145 - 9.0469x_1 - 33.421x_2 + 0.243x_3 - 0.5963x_1z_2 - 0.394x_1x_3 + 0.60022x_2x_3 + 0.6901x_1^2 + 11.7244x_2^2 - 0.0097x_3^2$$

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.

$$\hat{y}_{x_1=3,x_2=1.5,x_3=20} = 56.145 - 9.0469 \cdot 3 - 33.421 \cdot 1.5 + 0.243 \cdot 20 - 0.5963 \cdot 3 \cdot 1.5 - 0.0394 \cdot 3 \cdot 20 + 0.60022 \cdot 1.5 \cdot 20 +$$

$$0.6901 \cdot 3^2 + 11.7244 \cdot (1.5)^2 - 0.0097 \cdot 20^2 = 25.4028$$

b) ¿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?

$x_3 \longrightarrow x_3 + 1$

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

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

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

Ajuste bueno.

7) 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	1.5	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 $\hat{\theta} = (350.994, -1.27999, -015390)'$ y los valores ajustados son

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

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

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

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

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

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

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

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

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

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

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

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

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)

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

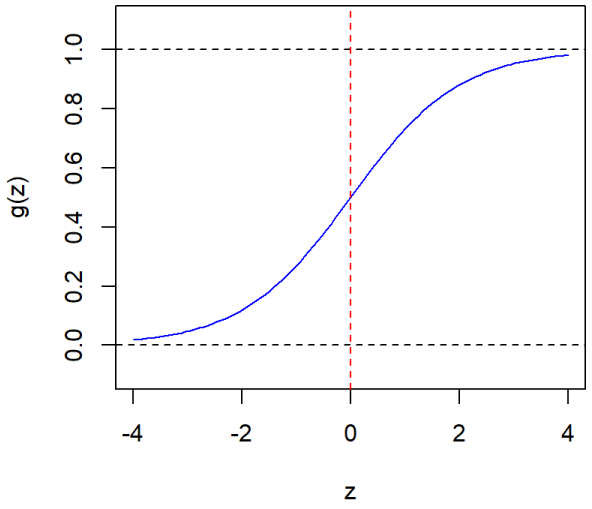
3.2) Función logística

```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

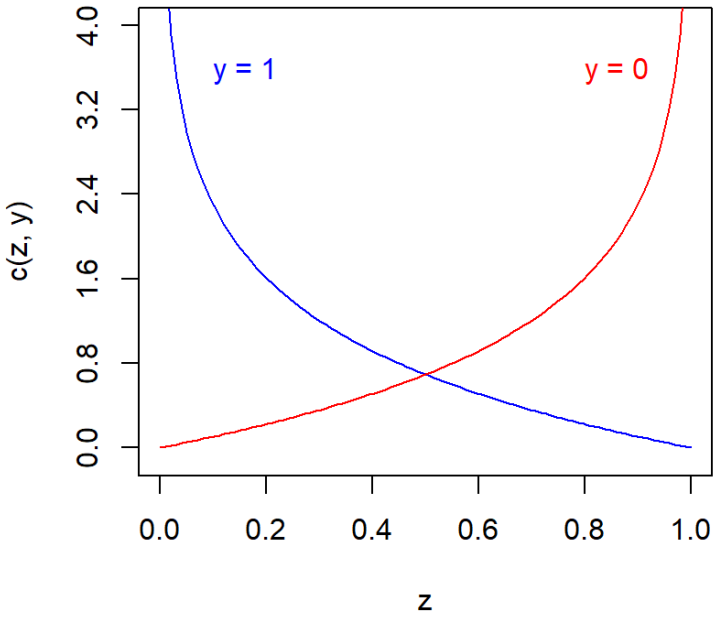
$$c(z, y) = \begin{cases} -\log(z) & \text{si } y = 1 \\ -\log(1 - z) & \text{si } y = 0 \end{cases}$$

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.

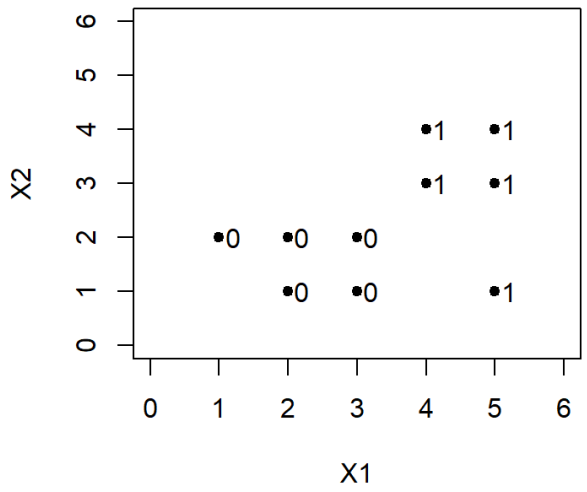
Supongamos que tenemos dos variables predictoras X_1 y X_2 ($k = 2$) y los datos siguientes:

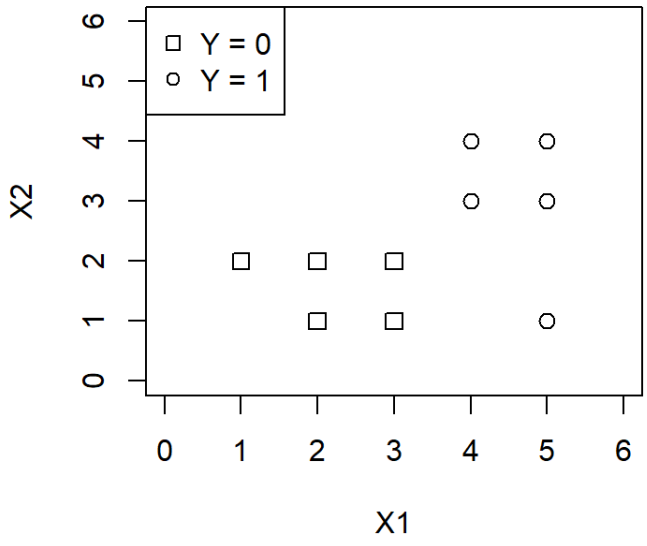
Individuo	X_1	X_2	Y
1	1	2	0
2	2	1	0
3	3	1	0
4	2	2	0
5	5	1	1
6	5	3	1
7	3	2	0
8	4	3	1
9	4	4	1
10	5	4	1

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:

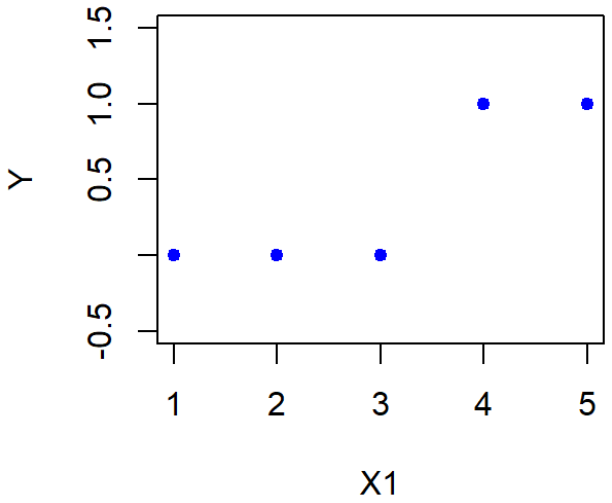
```r
X1 <- c(1, 2, 3, 2, 5, 5, 3, 4, 4, 5)
X2 <- c(2, 1, 1, 2, 1, 3, 2, 3, 4, 4)
Y <- c(0, 0, 0, 0, 1, 1, 0, 1, 1, 1)
tmp1 = tapply(X1, Y, mean)
tmp2 = tapply(X2, Y, mean)
tmp = rbind(tmp1, tmp2)
colnames(tmp) = paste0("Y = ", colnames(tmp))
rownames(tmp) = paste0("Media de ", c("X1", "X2"))
tmp
```

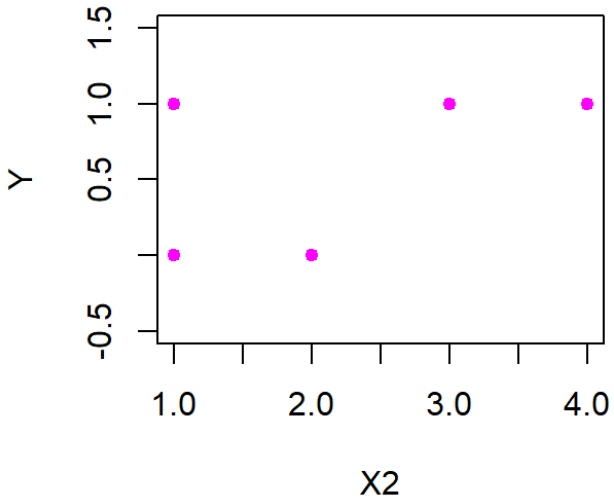
```
##             Y = 0 Y = 1
## Media de X1   2.2   4.6
## Media de X2   1.6   3.0
```

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.

```
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).

```
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

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

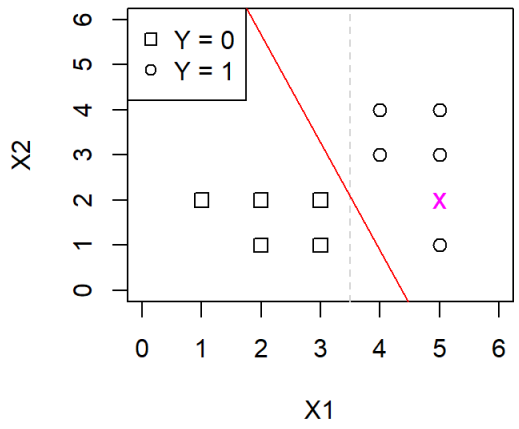
Esto es, $x_2 = 10,4506 - 2.3908x_1$

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.

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

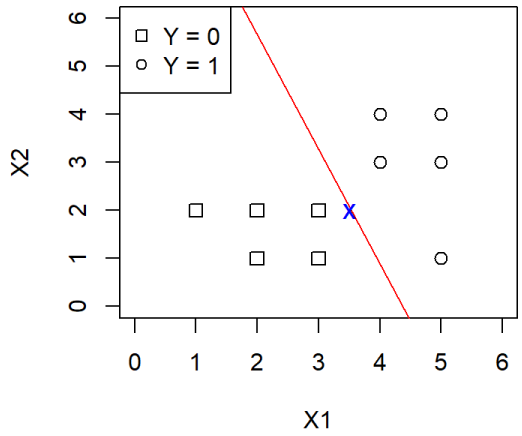
$Pr(Y = 0|X_1 = 5, X_2 = 2) \approx 1 - g(I_{\hat{\theta}}(5, 2)) = 0.0265$

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

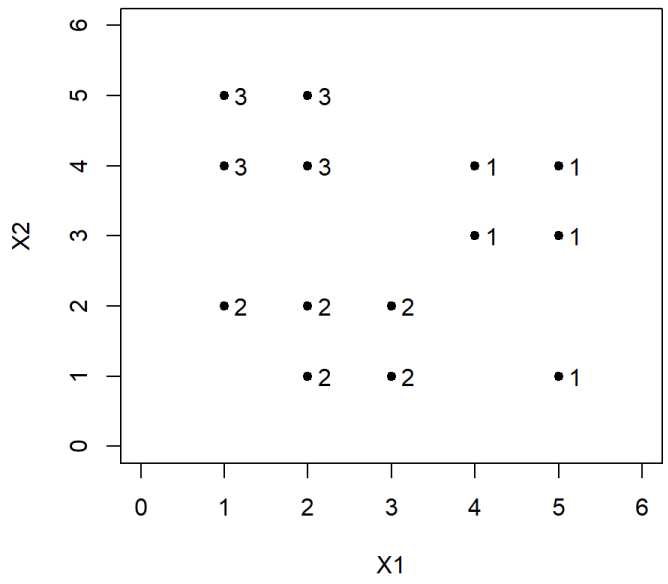
3.9.1) Cálculo de la verosimilitud

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)
```

```r
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.

```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.

```r
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

$$\theta = \begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 2 & 3 \end{pmatrix}$$

```r
theta = matrix(c(1, 2,
                 0, 1,
                 2, 3), nrow = k+1, ncol = g-1, byrow = TRUE)
J(theta)
```

```
## [1] 111.0598
```

Evaluamos en

$$\theta = \begin{pmatrix} -2.3 & 1.5 \\ 0.5 & 2 \\ 2.1 & 3.5 \end{pmatrix}$$

```r
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.

```r
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
```

```r
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.

RELACIÓN DE PROBLEMAS: REGRESIÓN LOGÍSTICA Y MULTINOMIAL
ANÁLISIS ESTADÍSTICO MULTIVARIANTE
GRADO EN CIENCIA E INGENIERÍA DE DATOS

1. En la siguiente tabla se ha recopilado una serie de 20 datos que relacionan las horas de estudio de cada alumno y si han aprobado o suspendido un examen de estadística.

Horas de estudio	Aprobado (1 sí, 0 no)	Horas de estudio	Aprobado (1 sí, 0 no)
0.5	0	2.75	1
0.75	0	3	0
1	0	3.25	1
1.25	0	3.5	0
1.5	0	4	1
1.75	0	4.25	1
1.75	1	4.5	1
2	0	4.75	1
2.25	1	5	1
2.5	0	5.5	1

Se ha ajustado un modelo de regresión logística y los parámetros estimados han sido $\hat{\theta}_0 = -4.077$ y $\hat{\theta}_1 = 1.5046$.

   a) ¿Cómo se interpreta el valor de $\hat{\theta}_1$?

   b) A partir del modelo ajustado, obtener una predicción para la probabilidad de que un alumno apruebe si ha estudiado 2 horas. ¿Cuál sería dicha probabilidad si dedicara una hora más al estudio? ¿Cómo ha variado la razón de estas probabilidades?

2. En un estudio clínico se desea predecir la probabilidad de padecer una enfermedad coronaria (Y, con valores 1 sí, 0 no) a partir de las covariables siguientes: Nivel de colesterol (X_1, 1 alto, 0 bajo), Edad (X_2) y Resultado del electrocardiograma (X_3, 1 anormal, 0 normal). Para ello, se analizaron 750 casos y se propuso un modelo logístico para estimar el riesgo de padecer una enfermedad coronaria, obteniendo las siguientes estimaciones para los parámetros: $\hat{\theta}_0 = -3.912$, $\hat{\theta}_1 = 0.852$, $\hat{\theta}_2 = 0.025$ y $\hat{\theta}_3 = 0.441$.

   a) Interpretar el significado de los coeficientes $\hat{\theta}_1$, $\hat{\theta}_2$ y $\hat{\theta}_3$.

   b) Obtener una predicción para la probabilidad de padecer una enfermedad coronaria para una persona con 40 años, electrocardiograma normal y nivel de colesteroal bajo. ¿Cuál sería dicha probabilidad si tuviera un nivel de colesterol alto?

   c) Para una persona con 40 años y electrocardiograma normal, ¿cómo influye el nivel de colesterol en el riesgo de padecer una enfermedad coronaria?

3. Se desea evaluar la satisfacción con la enseñanza pública de 1,500 estudiantes mediante la variable *Satisfecho* (Y, con valores 1 sí, 0 no) y tres variables predictoras: Nacionalidad (España=1, Ecuador = 2, Colombia=3), Género (Hombre=0, Mujer=1) y Estudios (ESO=0, Primaria=1). Se ajusta el siguiente modelo logístico:

$$\log \frac{p}{1-p} = -0.877 - 0.052\text{Nacionalidad2} + 1.72\text{Nacionalidad3} + 0.256\text{Género} - 0.008\text{Estudios},$$

donde $p = Pr[Y = 1]$. Las variables Nacionalidad2 y Nacionalidad3 son variables dicotómicas ficticias (*dummy*) que toman el valor 1 si el valor de Nacionalidad se corresponde con su índice y valen cero en caso contrario. Por ejemplo, si Nacionalidad = 3, entonces Nacionalidad2 = 0 y Nacionalidad3 = 1.

   a) Predecir la probabilidad de que una alumna colombiana de primaria no esté satisfecha con la enseñanza pública. ¿Cuál sería dicha probabilidad si la alumna tuviera nacionalidad española y estudiara primaria? ¿Y si fuera un alumno de secundaria con nacionalidad española?

   b) Comparar el grado de satisfacción con la enseñanza pública de alumnos de primaria con nacionalidad española según el género.

4. Una determinada compañía desea mejorar el marketing de cinco variedades de cereales para el desayuno. Para ello planifica un estudio encuestando a 900 personas, registrando su edad, género y si tiene o no un estilo de vida activo. Cada participante degustó los cinco tipos de cereales y se le preguntó sobre su preferencia. En ta tabla adjunta se presentan las definiciones de las variables.

Variable	Categoría
Tipo de cereal preferido	1: Cebada
	2: Centeno
	3: Avena
	4: Esbelta
	5: Trigo
Edad	1: Menor de 30 años
	2: 31 a 50 años
	3: Más de 50 años
Género	1: Hombre
	2: Mujer
Estilo de vida	0: No activo
(realiza o no actividad física)	1: Activo

   a) Con el objetivo de explicar la preferencia del tipo de cereal en función de la edad, el género y el estilo de vida se desea ajustar un modelo logístico multinomial. ¿Qué variables ficticias (*dummy*) debemos crear para la formulación del modelo?

   b) Especificar el modelo tomando el trigo como la categoría de referencia para la variable respuesta ($g = 5$).

   c) ¿Cómo se interpretan las estimaciones de los coeficientes del modelo $\hat{\theta}_{ij}$ que verifican que $\exp(\hat{\theta}_{ij}) > 1$? ¿Y si $\exp(\hat{\theta}_{ij}) < 1$?

1) En la siguiente tabla se ha recopilado una serie de 20 datos que relacionan las horas de estudio de cada alumno y si han aprobado o suspendido un examen de estadística.

Horas de estudio	Aprobado (1 sí, 0 no)	Horas de estudio	Aprobado (1 sí, 0 no)
0.5	0	2.75	1
0.75	0	3	0
1	0	3.25	1
1.25	0	3.5	0
1.5	0	4	1
1.75	0	4.25	1
1.75	1	4.5	1
2	0	4.75	1
2.25	1	5	1
2.5	0	5.5	1

Se ha ajustado un modelo de regresión y los parámetros estimados han sido $\hat{\theta}_0 = -4.077$ y $\hat{\theta}_1 = 1.5046$

a) ¿Como se interpreta el valor de $\hat{\theta}_1$?

$$\log \frac{p}{1-p} = \hat{\theta}_0 + \hat{\theta}_1 \cdot x$$

$$P \equiv Pr[Y = 1]$$

$$\frac{p}{1-p} = \exp(\hat{\theta}_0 + \hat{\theta}_1 x)$$

$$p = (1-p)\exp(\hat{\theta}_0 + \hat{\theta}_1 x)$$

b) A partir del modelo ajustado, obtener una predicción para la probabilidad de que un alumno apruebe si ha estudiado 2 horas. ¿Cuál sería dicha probabilidad si dedicara una hora más al estudio? ¿Cómo ha variado la razón de estas probabilidades?

$$Pr[Y = 1/X = 2] = \frac{1}{1 + \exp(-(-40.77 + 1.5046 \cdot 2))} \simeq 0.25$$

$$Pr[Y = 1/X = 3] = \frac{1}{1 + \exp(-(-40.77 + 1.5046 \cdot 3))} \simeq 0.607$$

$$\frac{Pr[Y = 1/X = 3]}{Pr[Y = 1/X = 2]} = \frac{0.607}{0.2556} = 2.428$$

La razón de ambas propiedades presenta una gran diferencia.

2) En un estudio clínico se desea predecir la probabilidad de padecer una enfermedad coronaria (Y, con valores 1 sí, 0 no) a partir de las covarianzas siguientes: Nivel de colesterol (X_1, 1 alto, 0 bajo), Edad (X_2) y Resultado del electrocardiograma (X_3, 1 anormal, 0 normal). Para ello, se analizaron 750 casos y se propuso un modelo logístico para estimar el riesgo de padecer una enfermedad coronaria, obteniendo las siguientes estimaciones para los parámetros: $\hat{\theta}_0 = -3.912, \hat{\theta}_1 = 0.852, \hat{\theta} = 0.025$ y $\hat{\theta}_3 = 0.441$.

a) Interpretar el significado de los coeficientes $\hat{\theta}_1, \hat{\theta}_2$ y $\hat{\theta}_3$.

$$\log\left(\frac{p}{1-p}\right) = -3.912 + 0.852 \cdot x_1 + 0.025 \cdot x_2 + 0.441 \cdot x_3$$

$$p = Pr[Y = 1]$$

$$\frac{p}{1-p} = \exp(-3.912 + 0.852 \cdot x_1 + 0.025 \cdot x_2 + 0.441 \cdot x_3) \longrightarrow p = (1-p)\exp(-3.912 + 0.852 \cdot x_1 + 0.025 \cdot x_2 + 0.441 \cdot x_3)$$

b) Obtener una predicción para la probabilidad de padecer una enfermedad coronaria para una persona con 40 años, electrocardiograma normal y nivel de colesterol bajo. ¿Cuál sería dicha probabilidad si tuviera un nivel de colesterol alto?

$$Pr(Y = 1/x_1 = 0, x_2 = 40, x_3 = 0) = \frac{1}{1 + e^{-(-3.912 + 0.025 \cdot 40)}} = 0.0516$$

$$Pr(Y = 1/x_1 = 1, x_2 = 40, x_3 = 0) = \frac{1}{1 + e^{-(-3.912 + 0.852 \cdot 1 + 0.025 \cdot 40)}} = 0.113$$

c) Para una persona con 40 años y electrocardiograma normal, ¿cómo influye el nivel de colesterol en el riesgo de padecer una enfermedad coronaria?

$$\frac{Pr(Y = 1/x_1 = 1, x_2 = 40, x_3 = 0)}{Pr(Y = 1/x_1 = 0, x_2 = 40, x_3 = 0)} = \frac{0.113}{0.0516} = 2.19$$

3) Se desea evaluar la satisfacción con la enseñanza pública de 1500 estudiantes mediante la variable *Satisfecho*(Y, con valores 1 sí, 0 no) y tres variables periódicas: Nacionalidad (España=1, Ecuador=2, Colombia=3), Género (Hombre=0, Mujer=1) y Estudios (ESO=0, Primaria=1). Se ajusta al siguiente modelo logístico:

$$\log \frac{p}{1-p} = -0.877 - 0.052 \cdot \text{Nacionalidad2} + 1.72 \cdot \text{Nacionalidad3} + 0.256 \cdot \text{Género} - 0.008 \cdot \text{Estudios}$$

donde $p = Pr[Y = 1]$. Las variables *Nacionalidad2* y *Nacionalidad3* son variables dicotómicas ficticias (*dummy*) que toman el valor 1 si el valor de nacionalidad se corresponde con su índice y valen cero en caso contrario. Por ejemplo, Si Nacionalidad = 3, entonces Nacionalidad2 = 0 y Nacionalidad3 = 1.

a) Predecir la probabilidad de que una alumna colombiana de primaria no esté satisfecha con la enseñanza pública. ¿Cuál sería dicha probabilidad si la alumna tuviera nacionalidad española y estudiara primaria? ¿Y si fuera un alumno de secundaria con nacionalidad española?

$$Pr[Y = 1|N2 = 0, N3 = 1, G = 1, E = 1] = \frac{1}{1 + \exp(-(-0.877 - 0.052 \cdot 0 + 1.72 \cdot 1 + 0.256 \cdot 1 - 0.008 \cdot 1))} = 0.7485$$
$$Pr[Y = 0|N2 = 0, N3 = 1, G = 1, E = 1] = 1 - 0.7485 = 0.2514$$
$$Pr[Y = 1|N2 = 1, N3 = 0, G = 1, E = 0] = \frac{1}{\exp(-(-0.877 - 0.052 \cdot 1 + 1.72 \cdot 0 + 0.256 \cdot 1 - 0.008 \cdot 0)} = 0.2937$$
$$Pr[Y = 0|N2 = 1, N3 = 0, G = 1, E = 0] = 1 - 0.2937 = 0.7063$$

b) Comparar el grado de satisfacción con la enseñanza pública de los alumnos de primaria con nacionalidad española según el género.

$$Pr[Y = 1|N2 = 1, N3 = 0, G = 0, E = 1] = \frac{1}{\exp(-(-0.877 - 0.052 \cdot 1 + 1.72 \cdot 0 + 0.256 \cdot 0 - 0.008 \cdot 1)} = 0.2921$$
$$Pr[Y = 1|N2 = 1, N3 = 0, G = 1, E = 1] = \frac{1}{\exp(-(-0.877 - 0.052 \cdot 1 + 1.72 \cdot 0 + 0.256 \cdot 1 - 0.008 \cdot 1)} = 0.3477$$
$$\frac{Pr[Y = 1|N2 = 1, N3 = 0, G = 1, E = 1]}{Pr[Y = 1|N2 = 1, N3 = 0, G = 0, E = 1]} = \frac{0.3477}{0.2921} = 1.19$$

4) Una determinada compañía desea mejorar el marketing de cinco variedades de cereales para el desayuno. Para ello planifica un estudio encuestando a 900 personas, registrando su edad, género y si tiene o no un estilo de vida activo. Cada participante degustó los cinco tipos de cereales y se le preguntó sobre su preferencia. En la tabla adjunta se presenta las definiciones de las variables

Variable	Categoría
Tipo de cereal preferido	1: Cebada
	2: Centeno
	3: Avena
	4: Esbelta
	5: Trigo
Edad	1: Menor de 30 años
	2: 31 a 50 años
	3: Más de 50 años
Género	1: Hombre
	2: Mujer
Estilo de vida	0: No activo
(realiza o no actividad física)	1: Activo

a) Con el objetivo de explicar la preferencia del tipo de cereal en función de la edad, el género y el estilo de vida se desea ajustar un modelo logístico multinomial. ¿Qué variables ficticias (*dummy*) debemos crear para la formulación del modelo?

Debemos crear dos variables ficticias para la edad:

$$Edad2 = \begin{cases} 1 & \text{si } Edad = 2 \text{ (31 a 50 años)} \\ 0 & \text{en caso contrario} \end{cases}$$

$$Edad3 = \begin{cases} 1 & \text{si } Edad = 3 \text{ (Más de 50 años)} \\ 0 & \text{en caso contrario} \end{cases}$$

b) Especificar el modelo tomando el trigo como la categoría de referencia para la variable respuesta ($g = 5$).

$Pj = Pr[Y = j]$, $j = 1, \ldots, 5$

$Y \equiv$ tipo de cereal preferido

$$\log \frac{p_j}{p_s} = \hat{\theta}_{j0} + \hat{\theta}_{j1} \cdot Edad2 + \hat{\theta}_{j2} \cdot Edad3 + \hat{\theta}_{j4} \cdot Estilo, \quad j = 1, \ldots, 4$$

c) ¿Cómo se interpretan las estimaciones de los coeficientes del modelo $\hat{\theta}_{ij}$ que verifican que $\exp(\hat{\theta}_{ij}) > 1$? ¿Y si $\exp(\hat{\theta}_{ij}) < 1$?

$\exp(\hat{\theta}_{ij}) > 1$ es equivalente a decir $\hat{\theta}_{ij} > 0$

$\exp(\hat{\theta}_{ij}) < 1$ es equivalente a decir $\hat{\theta}_{ij} < 0$

$\exp(\hat{\theta}_{ij}) = 1, \hat{\theta}_{ij} = 0 \Longrightarrow$ La variable b que acompaña el coeficiente no afecta en la predicción.

Tema 4: Análisis de componentes principales

4.1) Introducción

1) Objetivo

Simplificar la representación de datos multidimensionales al transformarlos en un nuevo conjunto de variables llamadas componentes principales.

- Reducción de dimensionalidad: a veces dispondremos de muchas variables y simplemente se querrá disminuir el número de variables perdiendo la menor información posible (compresión de datos).
  - Útil en aplicaciones de almacenamiento y transmisión de información.
- Visualización de datos: al proyectar los datos en un espacio de menor dimensión, es más fácil representar gráficamente la estructura subyacente de los datos, lo que puede ayudar a identificar patrones, agrupaciones o relaciones (extracción de características).
- Eliminación de multicolinealidad: en análisis de regresión y otros contextos, la multicolinealidad (alta correlación entre variables independientes) puede ser problemática.
  - En estos casos puede ayudar a reducir la multicolinealidad el transformar las variables originales en un conjunto de variables no correlacionadas (las componentes principales).

- Planteamiento desde el punto de vista teórico

La idea es resumir la información de un vector aleatorio (v.a.) k-dimensional $\mathbf{X} = (X_1, \ldots, X_k)'$ (recordemos que A' denota la traspuesta de A, es decir, $\mathbf{X}$ es un vector columna) en unas pocas variables que proporcionen la información más relevante.

Se puede dar una aproximación geométrica mediante el concepto de elipsoide de concentración.

- Definición

Si $\mathbf{X}$ es un vector aleatorio de dimensión k, media μ y su matriz de covarianzas $V = (\sigma_{i,j})$ definida positiva, se define el elipsoide de concentración de $\mathbf{X}$ como

$$E_k = \{\mathbf{x} \in \mathbb{R}^k : (\mathbf{x} - \mu)' V^{-1} (\mathbf{x} - \mu) \leq k + 2\}.$$

En la definición del elipsoide interviene la distancia de Mahalanobis basada en la matriz V entre $\mathbf{x}$ y la media μ dada por

$$d_V(\mathbf{x}, \mu) = \sqrt{(\mathbf{x} - \mu)' V^{-1} (\mathbf{x} - \mu)}.$$

Esta distancia al cuadrado se puede calcular en `R` con `mahalanobis(x, mu, V)`.

Además, si X es normal, el elipsoide se puede definir a partir de las curvas de nivel de la función de densidad ($f(x) = cte.$) ya que

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

Una parte de los individuos (puntos estarán dentro de este elipsoide).

Si queremos distinguirlos con una única variable, parece claro que lo mejor sería proyectarlos sobre el eje mayor del elipsoide.

Por ejemplo, para una normal bivariante

$$\mathcal{N}_2\left(\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, V = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}\right)$$

se tiene que

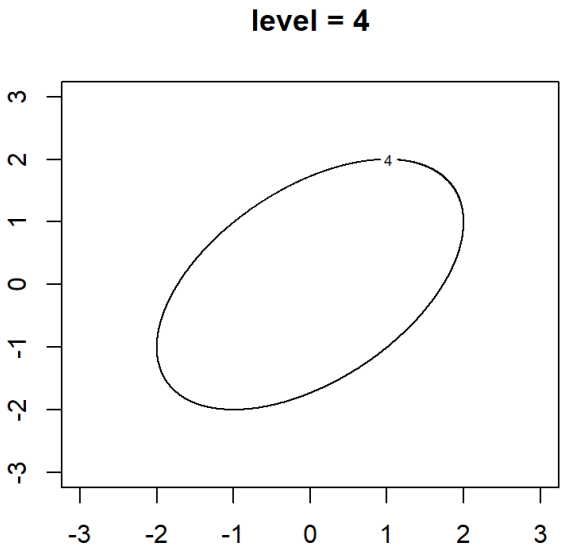
$$(x_1, x_2) \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}^{-1} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{4}{3}x_1^2 - \frac{4}{3}x_1 x_2 + \frac{4}{3}x_2^2$$

por lo que el elipsoide de concentración sería

$$\frac{4}{3}x_1^2 - \frac{4}{3}x_1 x_2 + \frac{4}{3}x_2^2 \leq 4$$

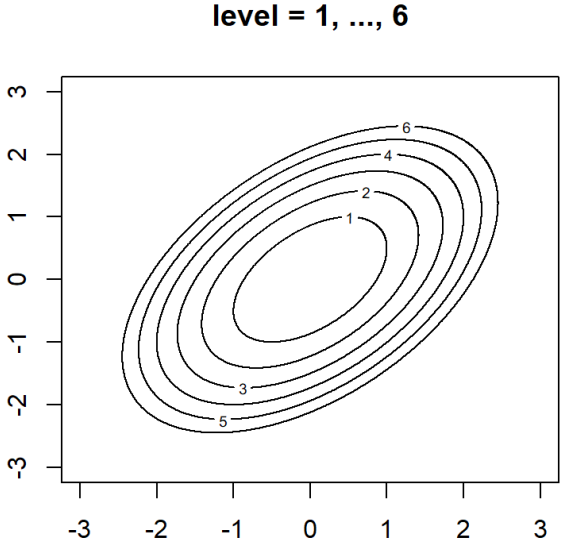
- Elipsoide de concentración para la normal bivariante con medias 0, varianzas 1 y correlación $\frac{1}{2}$:

```
hc <- function(x1, x2) (4/3)*x1^ 2- (4/3)*x1*x2 + (4/3)*x2^ 2
x1 <- seq(-3, 3, length =1000)
x2 <- seq(-3, 3, length =1000)
z <- outer(x1, x2, hc)
contour(x1, x2, z, levels = 4)
title(main = "level = 4")
```



- Elipsoides obtenidos con otros niveles (circunferencias de Mahalanobis):

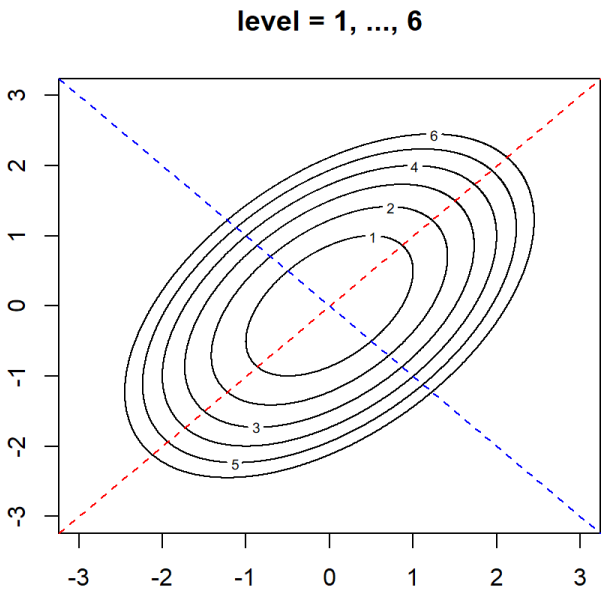
```
hc <- function(x1, x2) (4/3)*x1^ 2- (4/3)*x1*x2 + (4/3)*x2^ 2
x1 <- seq(-3, 3, length =1000)
x2 <- seq(-3, 3, length =1000)
z <- outer(x1, x2, hc)
contour(x1, x2, z, levels = c(1:6))
title(main = "level = 1, ..., 6")
```



Si queremos reducir las dos variables a solo una, la mejor proyección, es decir la que mejor separa los puntos (varianza máxima), es la proporcionada por el eje principal del elipsoide (o curvas de nivel de la normal).
En este ejemplo viene dado por la recta:

$$x_2 = x_1.$$

```r
hc <- function(x1, x2) (4/3)*x1^ 2- (4/3)*x1*x2
     + (4/3)*x2^ 2
x1 <- seq(-3, 3, length =1000)
x2 <- seq(-3, 3, length =1000)
z <- outer(x1, x2, hc)
contour(x1, x2, z, levels = c(1:6))
abline(a = 0, b = 1, col = "red", lty = 2)
abline(a = 0, b = -1, col = "blue", lty = 2)
title(main = "level = 1, ..., 6")
```



- Objetivo

Transformar un conjunto de k variables interrelacionadas entre sí en un nuevo conjunto con un número menor de variables:

- las componentes principales

De manera que estas nuevas variables:

- sean ortogonales entre sí.
- capturen la mayor variabilidad de las variables.
- expliquen la mayor parte de la variabilidad de las variables originales.

- Planteamiento teórico

Supongamos que $\mathbf{X} = (X_1, \ldots, X_k)'$ es un vector aleatorio k-dimensional con vector de medias μ y matriz de covarianzas V semidefinida positiva.

Entonces la primera componente principal será la variable aleatoria unidimensional

$$Y_1 = a_1 X_1 + \cdots + a_k X_k$$

con $a_1^2 + \cdots + a_k^2 = 1$ cuya varianza es máxima.

- Si no se normaliza la combinación lineal, la variable Y_1 puede tener varianza tan grande como queramos.

- Geométricamente, hacemos un cambio de variable (primer eje) para que la dispersión sea máxima y la normalización equivale a mantener la escala original (proyectar).

El problema puede expresarse de la forma siguiente:

$$\left.\begin{array}{ll} \max & \mathrm{Var}(a'\mathbf{X}) \\ \text{s.a.} & \mathbf{a}'\mathbf{a} = 1 \end{array}\right\}$$

donde $\mathbf{a} = (a_1, \ldots, a_k)' \in \mathbb{R}^k$.

Una vez calculada una primera componente principal Y_1, la segunda componente principal Y_2 debe verificar $\mathrm{Cov}(Y_1, Y_2) = 0$ (no debe contener información ya incluida en Y_1) y debe tener la varianza máxima, es decir,

$$\left.\begin{array}{ll} \max & \mathrm{Var}(\mathbf{a}'\mathbf{X}) \\ \text{s.a.} & \mathbf{a}'\mathbf{a} = 1 \\ & \mathrm{Cov}(Y_1, \mathbf{a}'\mathbf{X}) = 0 \end{array}\right\}$$

Así, sucesivamente, por inducción, se definen las siguientes componentes principales (Y_j) como la (una) solución de

$$\left.\begin{array}{ll} \max & \mathrm{Var}(\mathbf{a}'\mathbf{X}) \\ \text{s.a.} & \mathbf{a}'\mathbf{a} = 1 \\ & \mathrm{Cov}(Y_1, \mathbf{a}'\mathbf{X}) = 0, \quad i = 1, \ldots, j-1 \end{array}\right\}$$

- La solución general viene dada en el teorema siguiente que prueba la existencia de las (unas) componentes principales y muestra cómo calcularlas.
- Además, se demuestra que las componentes principales no son únicas (puede haber más soluciones).

- Teorema de existencia

Si $\mathbf{X}$ es un vector aleatorio k-dimensional con matriz de covarianzas V definida positiva, las (unas) componentes principales se obtienen como

$$\mathbf{Y} = (Y_1, \ldots, Y_k)' = T'\mathbf{X} = \begin{pmatrix} t_{1,1} & \cdots & t_{k,1} \\ \cdots & \cdots & \cdots \\ t_{1,k} & \cdots & t_{k,k} \end{pmatrix} \begin{pmatrix} X_1 \\ \cdots \\ X_k \end{pmatrix},$$

donde T es una matriz ortogonal $(T'T = TT' = I)$ tal que

$$T'VT = D = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$$

con $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > 0$.

- Demostración

Como V es una matriz simétrica y definida positiva, existe una matriz $T = (t_{i,j})$ ortogonal $(T'T = TT' = I)$ tal que

$$T'VT = D = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$$

con los valores propios verificando $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > 0$.

De esta forma, si

$$\mathbf{Y} = (Y_1, \ldots, Y_k)' = T'\mathbf{X} = \begin{pmatrix} t_{1,1} & \cdots & t_{k,1} \\ \cdots & \cdots & \cdots \\ t_{1,k} & \cdots & t_{k,k} \end{pmatrix} \begin{pmatrix} X_1 \\ \cdots \\ X_k \end{pmatrix}$$

entonces $Y_1, \ldots, Y_k$ verifican que

$$\mathrm{Cov}(\mathbf{Y}) = \mathrm{Cov}(T'\mathbf{X}) = E[T'(\mathbf{X} - \mu)(\mathbf{X} - \mu)'T] = T'VT = D,$$

lo que indica que $\mathrm{Cov}(Y_i, Y_j) = 0$ para $i \neq j$ y $\mathrm{Var}(Y_j) = \lambda_j$.

Para comprobar que Y_1 es una primera componente principal, supongamos que $\mathbf{a}'\mathbf{X}$ es una combinación lineal con $\mathbf{a}'\mathbf{a} = 1$.

Las columnas de la matriz T corresponden a los vectores propios $\mathbf{t}_i$ asociados a los autovalores $\lambda_i, T = (\mathbf{t}_1|\cdots|\mathbf{t}_k)$, y como los vectores propios son una base, existirán $c_1, \ldots, c_k$ números reales tales que

$$\mathbf{a} = c_1\mathbf{t}_1 + \cdots + c_k\mathbf{t}_k = \sum_{i=1}^{k} c_i\mathbf{t}_i.$$

Con lo que

$$\begin{aligned}\operatorname{Var}(\mathbf{a}'\mathbf{X}) &= E[\mathbf{a}'(\mathbf{X}-\mu)(\mathbf{X}-\mu)'\mathbf{a}] = \mathbf{a}'\operatorname{Cov}(\mathbf{X})\mathbf{a} = \mathbf{a}'V\mathbf{a} \\ &= \left(\sum_{i=1}^{k} c_i\mathbf{t}_i'\right) V \left(\sum_{i=1}^{k} c_i\mathbf{t}_i\right) = \left(\sum_{i=1}^{k} c_i\mathbf{t}_i'\right)\left(\sum_{j=1}^{k} c_j V\mathbf{t}_j\right) \\ &= \left(\sum_{i=1}^{k} c_i\mathbf{t}_i'\right)\left(\sum_{j=1}^{k} c_j\lambda_j\mathbf{t}_j\right) = \sum_{i,j} c_ic_j\lambda_j\mathbf{t}_i'\mathbf{t}_j = \sum_{i=1}^{k} c_i^2\lambda_i\end{aligned}$$

Y, como

$$\mathbf{a}'\mathbf{a} = \left(\sum_{i=1}^{k} c_i\mathbf{t}_i'\right)\left(\sum_{j=1}^{k} c_j\mathbf{t}_j\right) = \sum_{i,j} c_ic_j\mathbf{t}_i'\mathbf{t}_j = \sum_{i=1}^{k} c_i^2 = \mathbf{c}'\mathbf{c} = 1,$$

con $\mathbf{c} = (c_1, \ldots, c_k)'$, la varianza será máxima si $x_1^2 = 1, c_2 = 0, \ldots, c_k = 0$ ya que

$$\operatorname{Var}(\pm\mathbf{t}_1'\mathbf{X}) = \lambda_1 = \mathbf{c}'\mathbf{c}\lambda_1 = \sum_{i=1}^{k} c_i^2\lambda_1 \geq \sum_{i=1}^{k} c_i^2\lambda_i = \operatorname{Var}(\mathbf{a}'\mathbf{X}),$$

para todo $\mathbf{a}$ tal que $\mathbf{a}'\mathbf{a} = 1$, es decir, $Y_1 = \pm\mathbf{t}_1'\mathbf{X}$ es una primera componente principal (puede haber otras soluciones si $\lambda_1 = \lambda_2$).

Por inducción, supongamos que $Y_1 = \mathbf{t}_1'\mathbf{X}, \ldots, Y_{j-1}\mathbf{X}$ son las primeras $(j-1)$ componentes principales.

Y veamos que $Y_j = \mathbf{t}_j'\mathbf{X}$ es la (una) solución de

$$\left.\begin{array}{ll} \max & \operatorname{Var}(\mathbf{a}'\mathbf{X}) \\ \text{s.a.} & \mathbf{a}'\mathbf{a} = 1 \\ & \operatorname{Cov}(Y_1, \mathbf{a}'\mathbf{X}) = 0, \quad i = 1, \ldots, j-1 \end{array}\right\}$$

Como se debe verificar

$$\begin{aligned}\operatorname{Cov}(\mathbf{a}'\mathbf{X}, Y_i) &= \operatorname{Cov}(\mathbf{a}'\mathbf{X}, \mathbf{t}_i'\mathbf{X}) = E[\mathbf{a}'(\mathbf{X}-\mu)(\mathbf{X}-\mu)'\mathbf{t}_i] = \mathbf{a}'\operatorname{Cov}(\mathbf{X})\mathbf{t}_i \\ &= \mathbf{a}'V\mathbf{t}_i = \mathbf{a}'\lambda_i\mathbf{t}_i = \lambda_i\mathbf{a}'\mathbf{t}_i = \lambda_i\left(\sum_s c_s\mathbf{t}_s'\right)\mathbf{t}_i = \lambda_ic_i = 0\end{aligned}$$

para $i = 1, \ldots, j-1$, $\lambda_i > 0$, se tiene $c_1 = \cdots = c_{j-1} = 0$.

Entonces, la varianza será máxima si $c_j = 1$ y $c_i = 0$ para $i > j$, ya que

$$\operatorname{Var}(\pm\mathbf{t}_j'\mathbf{X}) = \lambda_j = \mathbf{c}'\mathbf{c}\lambda_j = \sum_{i=j}^{k} c_i^2\lambda_j \geq \sum_{i=j}^{k} c_i^2\lambda_i = \operatorname{Var}(\mathbf{a}'\mathbf{X}),$$

para todo $\mathbf{a}$ tal que $\mathbf{a}'\mathbf{a} = 1$ y $\operatorname{Cov}(\mathbf{a}'\mathbf{X}, Y_i) = 0$, $i = 1, \ldots, j-1$, es decir, $Y_j = \pm\mathbf{t}_j'\mathbf{X}$ es una componente principal j-ésima (no necesariamente la única).

- Corolario

Si $\lambda_1 > \lambda_2 > \cdots > \lambda_k$, entonces las componentes principales son únicas salvo digno.

- Observación

Nótese que la componente principal j-ésima se obtiene multiplicando la fila j-ésima de T' (la columna j-ésima de T) por $\mathbf{X}$, es decir,

$$Y_j = \mathbf{t}_j'\mathbf{X}$$

donde $\mathbf{t}'_j = (t_{1,j}, \ldots, t_{k,j})$ es un vector propio unitario correspondiente al j-ésimo valor propio (vectores columna de T).

Además, $\mathrm{Var}(Y_j) = \lambda_j$, y

$$\mathrm{traza}(V) = \sum_{j=1}^{k} \sigma_{j,j} = \sum_{j=1}^{k} \mathrm{Var}(X_j) = \sum_{j=1}^{k} \mathrm{Var}(Y_j) = \sum_{j=1}^{k} \lambda_j$$

(las matrices semejantes tienen las trazas iguales), es decir, la variabilidad (información) de las variables originales es igual a la suma de las variabilidades de las componentes principales.

La cantidad de información (%) contenida en cada componente será

$$I_j = 100 \frac{\lambda_j}{\sum_{i=1}^{k} \lambda_i} \%.$$

Por esto, la traza se usar como una medida unidimensional de la dispersión de una variable k-dimensional.

La otra medida es el determinante de V para el que también se verifica:

$$|V| = \lambda_1 \cdots \lambda_k = |\operatorname{Cov}(\mathbf{Y})|$$

- Observación

Otros autores llaman componentes principales a

$$\mathbf{Y} = T'(\mathbf{X} - \mu)$$

con lo que, además, se consigue que sean centradas ($E[Y_j] = 0$),

- compronentes principales centradas

También se pueden definir las componentes principales estandarizadas

$$Z_j = \mathbf{t}'_j(\mathbf{X} - \mu)\lambda_j^{-\frac{1}{2}}$$

($\mathbf{Z} = D^{-\frac{1}{2}}T'(\mathbf{X} - \mu)$) que además de ser centradas tendrán varianza 1.

Cuando hay valores propios iguales a cero (V es semidefinida positiva) no suelen considerarse sus correspondientes componentes principales (degeneradas) y se puede conservar toda la información en las componentes principales de valores propios distintos de cero.

En este caso hay variables que pueden obtenerse como combinación lineal de las restantes (aunque no siempre pueden eliminarse del análisis).

Geométricamente, las componentes principales se corresponden con los ejes principales del elipsoide de concentración.

Como $\mathbf{Y} = T'\mathbf{X}$, podemos interpretar las componentes en función de los pesos que tengan en ellas las variables originales.

Si ponemos $\mathbf{X}$ en función de $\mathbf{Y}$ como $\mathbf{X} = T\mathbf{Y}$, entonces las variables originales se pueden interpretar en función de las componentes principales e incluso, podemos representar aproximadamente, las variables originales usando las dos (tres) primeras componentes.

- Caso de normalidad

Si la población $\mathbf{X}$ es normal, entonces las componentes principales son normales e independientes entre sí, ya que en estas poblaciones equivalen los conceptos de independencia e incorrelación (independencia lineal) y $\mathbf{Z}$ será una normal estándar multivariante ($\mathcal{N}_k(0, I)$).

- Proposición

Si $\mathbf{Y}$ son las componentes principales obtenidas a partir de $\mathbf{X}$, entonces $\mathbf{X}$ es normal multivariante si, y sólo si $Y_1, \ldots, Y_k$ son independientes y normales univariantes para todo $j = 1, \ldots, k$.

- Demostración.

La demostración es inmediata.

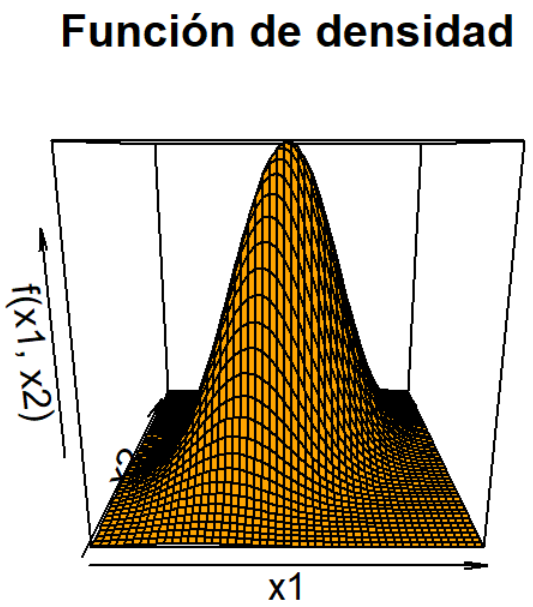
Esta propiedad puede ser utilizada para estudiar la normalidad multivariante a partir de un test de normalidad univariante sobre las componentes principales.

Incluso si la normal multivariante no es de rango completo (V no es definida positiva), puede utilizarse con con las m primeras componentes con valores propios distintos de cero (las otras serán degeneradas) coincidiendo m con el rango de V.

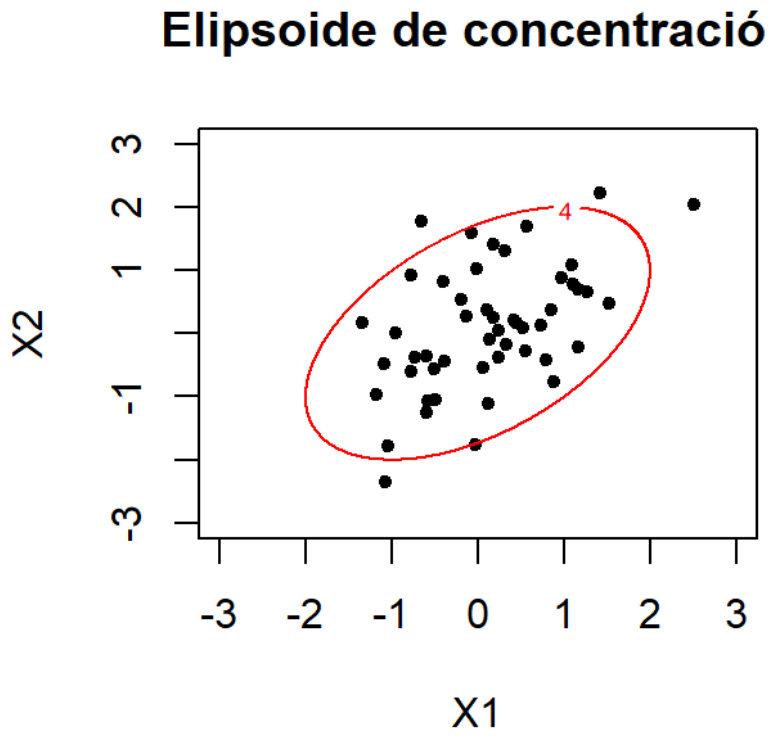
Ejemplo

Para el vector aleatorio normal de media $\mu = (0, 0)$ y matriz de covarianzas $V = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}$

```r
library("mvtnorm")
f <- function(x1, x2) dmvnorm(data.frame(x1, x2), mu, V)
V <- matrix(c(1, 1/2,
1/2, 1), nrow = 2, ncol = 2, byrow = TRUE)
mu <- c(0, 0)
x1 <- seq(-3, 3, length = 50)
x2 <- seq(-3, 3, length = 50)
z <- outer(x1, x2, f)
persp(x1, x2, z, xlab = 'x1', ylab = 'x2', zlab = 'f(x1, x2)', col = 'orange', main = "Función
    de densidad")
```



```r
#Se fija la semilla para la generación aleatoria
set.seed(123)
d <- rmvnorm(50, mu, V)
plot(d, xlab = "X1", ylab = "X2", pch = 20, xlim = c(-3, 3), ylim = c(-3, 3), main = "Elipsoide
    de concentración")
hc <- function(x1, x2) (4/3)*x1^ 2 - (4/3)*x1*x2 + (4/3)*x2^ 2
x1 <- seq(-3, 3, length = 1000)
x2 <- seq(-3, 3, length = 1000)
z <- outer(x1, x2, hc)
contour(x1, x2, z, levels = 4, add = T, col = 'red')
```



Sus componentes principales se calcularán diagonalizando V mediante

$$|V - \lambda I| = \begin{vmatrix} 1-\lambda & 0.5 \\ 0.5 & 1-\lambda \end{vmatrix} = 1 - 2\lambda + \lambda^2 - \frac{1}{4} = 0$$

que tiene soluciones

$$\lambda = \frac{2 \pm \sqrt{4 - 4(1 - \frac{1}{4})}}{2} = 1 \pm 0.5,$$

$\lambda_1 = 1.5$ y $\lambda_2 = 0.5$.

Y la primera componente se obtendrá resolviendo $V\mathbf{v} = \lambda\mathbf{v}$

$$\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 1.5 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \Longrightarrow \begin{pmatrix} -0.5x_1 + 0.5x_2 \\ 0.5x_1 - 0.5x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

lo que da $x_1 = x_2$, es decir, sus vectores propios son de la forma $\mathbf{v} = \alpha(1, 1)'$.

Como usamos vectores normalizados (de norma 1), una primera componente valdrá

$$Y_1 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \frac{X_1 + X_2}{\sqrt{2}}$$

y su varianza es $\lambda_1 = 1.5$.

Análogamente, la segunda valdrá $Y_2 = \dfrac{X_1 - X_2}{\sqrt{2}}$ (ya que tiene que ser perpendicular a la primera) y tendrá varianza $\lambda_2 = 0.5$.

Es decir, tenemos

$$Y_1 = \frac{X_1 + X_2}{\sqrt{2}}$$
$$Y_2 = \frac{X_1 - X_2}{\sqrt{2}}$$

por lo que

$$\mathbf{Y} = T'\mathbf{X} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}.$$

Varianza total explicada por cada componente:

- La primera componente explicará un 75% de la varianza total:

$$I_1 = 100\frac{\lambda_1}{\lambda_1 + \lambda_2}\% = 75\%.$$

- La segunda un 25% de la varianza total:

$$I_2 = 100\frac{\lambda_2}{\lambda_1 + \lambda_2}\% = 25\%.$$

Como las varianzas iniciales son iguales, ambas tienen igual peso en las componentes con distinto signo en el caso de la segunda de ellas.

Aunque las varianzas iniciales sean todas iguales (1) las componentes principales tienen varianzas (en general) distintas.

Si X_1 fuese el peso de una persona y X_2 su altura (estandarizadas).

- La primera componente se podría interpretar como lo grande que es dicha persona.
- Mientras que la segunda estará relacionada con su constitución (Y_2 grande significaría mucho peso y poca altura, es decir, complexión fuerte).

Despejando, se tiene

$$X_1 = \frac{Y_1 + Y_2}{\sqrt{2}}$$
$$X_2 = \frac{Y_1 - Y_2}{\sqrt{2}}$$

lo que nos permite representar las variables X_1, X_2 en función de las componentes Y_1, Y_2.

- Y_1 aumenta si lo hacen X_1 y X_2.
- Y_2 aumenta si aumenta X_1 y disminuye X_2.
- Estas relaciones servirán para interpretar (dar significado) a las componentes principales.

4.2) ¿Cómo realizamos estos cálculo en R?

En primer lugar definimos e introducimos V

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

```
##      [,1] [,2]
## [1,]  1.0  0.5
## [2,]  0.5  1.0
```

Calculamos los valores y vectores propios:

```
eigen(V)$values; eigen(V)$vectors
```

```
## [1] 1.5 0.5
##           [,1]       [,2]
## [1,] 0.7071068 -0.7071068
## [2,] 0.7071068  0.7071068
```

Podemos guardar la matriz T de vectores propios:

```
T <- eigen(V)$vectors
```

Los vectores normalizados aparecen en las columnas de T. Podemos comprobar que T es una matriz ortogonal:

```
t(T) %*% T
```

```
##      [,1] [,2]
## [1,]    1    0
## [2,]    0    1
```

donde `t(A)` es la traspuesta de `A` y `A %*% B` es el producto de las matrices `A` y `B` en `R`.

Podemos comprobar que T diagonaliza a V:

```
t(T) %*% V %*% T
```

```
##      [,1] [,2]
## [1,]  1.5  0.0
## [2,]  0.0  0.5
```

lo que nos dará la matriz diagonal con los valores 1.5 y 0.5 en la diagonal.

Como $\mathbf{Y} = T'\mathbf{X}$, las componentes principales serán

$$Y_1 = 0.7071068X_1 + 0.7071068X_2 = \frac{X_1 + X_2}{\sqrt{2}}$$
$$Y_2 = -0.7071068X_1 + 0.7071068X_2 = -\frac{X_1 + X_2}{\sqrt{2}}$$

Para calcular las informaciones contenidas en cada una (en tanto por 100) haremos:

```
100*eigen(V)$values/sum(eigen(V)$values)
```

```
## [1] 75 25
```

obteniendo el 75% y el 25%

4.3) Desigualdades

Si Z es una variable aleatoria no negativa con media finita $E[Z]$ y $\epsilon > 0$, entonces

$$\epsilon Pr[Z \geq \epsilon] = \epsilon \int_{[\epsilon,\infty)} \mathrm{d}F_Z(x) \leq \int_{[\epsilon,\infty)} x\mathrm{d}F_Z(x) \leq \int_{[0,\infty)} x\mathrm{d}F_Z(x) = E(Z)$$

(donde $F_Z(x) = Pr[Z \leq x]$ es su función de distribución), es decir

$$Pr[Z \geq \epsilon] \leq \frac{E[Z]}{\epsilon}.$$

$$E\left[\frac{(X-\mu)^2}{\sigma^2}\right] = \frac{1}{\sigma^2}E[(X-\mu)^2] = \frac{\sigma^2}{\sigma^2} = 1$$

Si X es una variable aleatoria con media finita $\mu = E[X]$ y varianza $\sigma^2 = \text{Var}(X) > 0$, entonces tomando $Z = \frac{(X-\mu)^2}{\sigma^2} \geq 0$ y aplicando la desigualdad de Markov, tenemos

$$Pr\left[\frac{(X-\mu)^2}{\sigma^2} \geq \epsilon\right] \leq \frac{1}{\epsilon}$$

para todo $\epsilon > 0$.

4.3.1) Desigualdad de Chebyshev

También se puede escribir como

$$Pr[(X-\mu)^2 < \epsilon\sigma^2] \geq 1 - \frac{1}{\epsilon},$$

o como

$$Pr[|X-\mu| < r] \geq 1 - \frac{\sigma^2}{r^2},$$

para todo $r > 0$.

4.3.2) Desigualdad de Chebyshev multivariante

Sea $\mathbf{X} = (X_1, \ldots, X_k)'$ un vector aleatorio con vector de medias finito $\mu = E(\mathbf{X})$ y matriz de covarianzas definida positiva V, entonces

$$Pr[(\mathbf{X}-\mu)'V^{-1}(\mathbf{X}-\mu) \geq \epsilon] \leq \frac{k}{\epsilon}$$

para todo $\epsilon > 0$.

- Consecuencias

La desigualdad también se puede escribir como

$$Pr[(\mathbf{X}-\mu)'V^{-1}(\mathbf{X}-\mu) < \epsilon] \geq 1 - \frac{k}{\epsilon},$$

para todo $\epsilon > 0$.

En particular, para el elipsoide de concentración

$$E_k = \{x \in \mathbb{R}^k : (\mathbf{X}-\mu)V^{-1}(\mathbf{X}-\mu) \leq k+2\},$$

obtenemos

$$Pr[\mathbf{X} \in E_k] \geq 1 - \frac{k}{k+2} = \frac{2}{k+2}.$$

Para obtener regiones con más datos podemos tomar $\epsilon = ck$, resultando

$$Pr[(\mathbf{X}-\mu)'V^{-1}(\mathbf{X}-\mu) < ck] \geq 1 - \frac{k}{\epsilon} = 1 - \frac{1}{c} = \frac{c-1}{c}.$$

La variable aleatoria no negativa $Z = (\mathbf{X}-\mu)'V^{-1}(\mathbf{X}-\mu)$ se puede escribir como

$$(\mathbf{X}-\mu)'TD^{-1}T'(\mathbf{X}-\mu) = \left[D^{-\frac{1}{2}}T'(\mathbf{X}-\mu)\right]'\left[D^{-\frac{1}{2}}T'(\mathbf{X}-\mu)\right] = \mathbf{Z}'\mathbf{Z},$$

donde $\mathbf{Z} = D^{-\frac{1}{2}}T'(\mathbf{X}-\mu)$ ($\mathbf{Z} = (Z_1, \ldots, Z_k)'$).

Si X es normal, entonces $Z_1, \ldots, Z_k$ son normales estándar independientes y

$$Z = \sum_{i=1}^{k} Z_i^2$$

sigue una distribución chi-cuadrado con k grados de libertad (ya que es la suma de k normales $\mathcal{N}(0,1)$ independientes).

4.4) Propiedades

- Proposición

Si $\mathbf{Y}$ son las componentes principales obtenidas a partir de $\mathbf{X}$, entonces

$$\begin{aligned}\text{Cov}(\mathbf{X},\mathbf{Y}) &= TD\\ \text{Corr}(\mathbf{X},\mathbf{Y}) &= \text{diag}(V)^{-\frac{1}{2}}TD^{\frac{1}{2}}\end{aligned}$$

donde $\text{diag}(V) = \text{diag}(\sigma_1^2,\ldots,\sigma_k^2)$.

- Demostración

En primer lugar señalaremos que

$$\text{Cov}(\mathbf{X},\mathbf{Y}) = \text{Cov}(\mathbf{X},T'\mathbf{X}) = VT$$

y, como $T'VT = D$ y T es ortogonal, entonces $VT = TD$ y $\text{Cov}(\mathbf{X},\mathbf{Y}) = TD$.

Por otro lado se tiene que como

$$\text{Corr}(X_i,Y_j) = \frac{\text{Cov}(X_i,Y_j)}{\sigma_i\lambda_j^{\frac{1}{2}}},$$

entonces

$$\text{Corr}(\mathbf{X},\mathbf{Y}) = \text{diag}(V)^{-\frac{1}{2}}\,\text{Cov}(\mathbf{X},\mathbf{Y})D^{-\frac{1}{2}}$$

y

$$\text{Corr}(\mathbf{X},\mathbf{Y}) = \text{diag}(V)^{-\frac{1}{2}}TDD^{-\frac{1}{2}} = \text{diag}(V)^{-\frac{1}{2}}TD^{\frac{1}{2}}$$

- Corolario

En las condiciones de la proposición anterior se tiene:

$$\begin{aligned}\text{Cov}(X_i,Y_j) &= t_{i,j}\lambda_j\\ \text{Corr}(X_i,Y_j) &= \frac{t_{i,j}}{\sigma_i}\lambda_j^{\frac{1}{2}}\end{aligned}$$

para todo i,j.

- Definición

Se denomina matriz de saturaciones a

$$A = \text{Corr}(\mathbf{X},\mathbf{Y}).$$

Ejemplo

Para el vector aleatorio normal de media $\mu = (0,0)$ y matriz de covarianzas $V = \begin{pmatrix}1 & \frac{1}{2}\\ \frac{1}{2} & 1\end{pmatrix}$, se obtiene

$$T = (\mathbf{t}_1|\mathbf{t}_2) = \begin{pmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}\end{pmatrix}, \quad D = \begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{pmatrix} = \begin{pmatrix}1.5 & 0\\ 0 & 0.5\end{pmatrix}$$

por lo que la matriz de saturaciones valdrá:

$$A = \text{diag}(V)^{-\frac{1}{2}}TD^{\frac{1}{2}} = \frac{1}{2}\begin{pmatrix}\sqrt{3} & 1\\ \sqrt{3} & -1\end{pmatrix} = \begin{pmatrix}0.86603 & 0.5\\ 0.86603 & -0.5\end{pmatrix}$$

Nótese que:

- La primera componente explica un 75% (0.866$\hat{}$ 2 $ 100) de las variables X_1 y X_2.

- Mientras que la segunda solo un 25%.

Las saturaciones y sus caudrados suelen representarse en tablas de la forma siguiente:

$a_{i,j} = \text{Corr}(X_i, Y_j)$	Y_1	Y_2
X_1	0.866	0.5
X_2	0.866	-0.5

$a_{i,j}$	Y_1	Y_2	Total
X_1	0.75	0.25	1
X_2	0.75	0.25	1

lo que nos puede ayudar a interceptar las componentes principales.

Las saturaciones también se pueden representar gráficamente.

Aunque en este ejemplo, las saturaciones con las distintas variables coincidan, esto no siempre es así, y tendremos variables mejor explicadas por las componentes elegidas que otras.

- Proposición

Si A es la matriz de saturaciones, entonces

$$AA' = \text{Corr}(\mathbf{X}).$$

También es interesante calcular las correlaciones múltiples entre cada variable original con el grupo de las p primeras componentes principales elegidas ($p \leq k$).

- Para medir el máximo que podemos explicar de cada variable original a partir de combinaciones lineales de esas componentes principales.

- Proposición

Si $\mathbf{Y}$ son las componentes principales obtenidas a partir de $\mathbf{X}$, entonces

$$\text{Corr}^2(X_i, (Y_1, \ldots, Y_p)) = \sum_{j=1}^{p} \text{Corr}^2(X_i, Y_j) = \frac{1}{\sigma_{i,i}} \sum_{j=1}^{p} t_{i,j}^2 \lambda_j = \sum_{j=1}^{p} a_{i,j}^2.$$

La demostración es inmediata ya que las componentes son incorreladas entre sí.

- Definición

A estas correlaciones se las suele denominar comunalidades

$$c_i = \text{Corr}^2(X_i, (Y_1, \ldots, Y_p))$$

y se suelen representar en la tabla de las saturaciones al cuadrado (como totales de las filas).

Además, el máximo de la correlación se obtiene con la combinación lineal $\alpha_i'(Y_1, \ldots, Y_p)'$ con

$$\alpha_i = \lambda V_{2,2}^{-1} v_{1,2} = \lambda \begin{pmatrix} \lambda_1^{-1} & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & \lambda_p^{-1} \end{pmatrix} \begin{pmatrix} t_{i,1}\lambda_1 \\ \cdots \\ t_{i,p}\lambda_p \end{pmatrix} = \lambda \begin{pmatrix} t_{i,1} \\ \cdots \\ t_{i,p} \end{pmatrix}.$$

Es decir, si tenemos que obtener $\mathbf{X}$ en función de las p primeras componentes principales, lo haremos a partir de la relación $\mathbf{X} = T\mathbf{Y}$ eliminando el resto de las componentes.

Lógicamente, si $p = k$, se obtiene $\alpha_i'(Y_1, \ldots, Y_p) = \lambda X_i$ y

$$\text{Corr}^2(X_i, (Y_1, \ldots, Y_k)) = \sum_{j=1}^{k} \text{Corr}^2(X_i, Y_j) = \frac{1}{\sigma_{i,i}} \sum_{j=1}^{k} t_{i,j}^2 \lambda_j = 1.$$

Recíprocamente, la información contenida en la componente principal j-ésima vale:

$$\lambda_j = \lambda_j \sum_{i=1}^{k} t_{i,j}^2 = \sum_{i=1}^{k} \sigma_{i,i} \frac{1}{\sigma_{i,i}} t_{i,j}^2 \lambda_j = \sum_{i=1}^{k} \sigma_{i,i} \operatorname{Corr}^2(X_i, Y_j),$$

ya que $\sum_{i=1}^{k} t_{i,j}^2 = 1$ es el módulo al cuadrado del vector propio $\mathbf{t}_j$ (columnas de T).

Y la información (variación) total contenida en las p primeras componentes principales vale:

$$\sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} \sum_{i=1}^{k} \sigma_{i,i} \operatorname{Corr}^2(X_i, Y_j) = \sum_{i=1}^{k} \sigma_{i,i} \operatorname{Corr}^2(X_i, (Y_1, \ldots, Y_p)) = \sum_{i=1}^{k} c_i \sigma_i^2.$$

Si todas las varianzas son 1, la información total $\sum_{j=1}^{p} \lambda_j$ será la suma de las comunalidades, es decir, la suma de la información que se tiene de cada variable original.

- Si $p = k$, entonces $c_i = 1$ y se tiene

$$\sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} \sigma_i^2.$$

- Seguimos con el ejemplo ...

En el ejemplo anterior obtuvimos que:

$a_{i,j}$	Y_1	Y_2	Total
X_1	0.75	0.25	1
X_2	0.75	0.25	1
Total	1.5	0.25	2

donde:

- Si $p = 1$, se tiene que $\lambda_1 = \frac{3}{2} = 0.75 + 0.75$.
- Si $p = 2$, se tiene que $\lambda_1 + \lambda_2 = \frac{3}{2} + \frac{1}{2} = 2 = 1 + 1 = \sigma_1^2 + \sigma_2^2$.

4.5) Cálculo a partir de la matriz de correlaciones

Cuando se estudian variables en las que se usan unidades diferentes o queremos que éstas no sean significativas (todas las variables sean iguales a priori), las componentes principales suelen calcularse a partir de la matriz de correlaciones

$$\Pi = (\rho_{i,j})$$

con $\rho_{i,j} = \frac{\sigma_{i,j}}{\sigma_i \sigma_j}$.

Esto equivale a considerar desde el principio las variables estandarizadas

$$Z_i = \frac{X_i - \mu_i}{\sigma_i}$$

(se igualan las varianzas a 1).

Usando el teorema principal, se obtienen las componentes

$$\tilde{\mathbf{Y}} = \tilde{T}' \mathbf{Z} = \tilde{T}' \operatorname{diag}(V)^{-\frac{1}{2}} (\mathbf{X} - \mu)$$
$$\tilde{Y}_j = \tilde{\mathbf{t}}_j' \mathbf{Z} = \sum_{i=1}^{k} \tilde{t}_{i,j} Z_i = \sum_{i=1}^{k} \tilde{t}_{i,j} \frac{X_i - \mu_i}{\sigma_i}$$

donde $\tilde{T}$ es la matriz ortogonal que diagonaliza $\Pi = \operatorname{Corr}(\mathbf{X}) = \operatorname{Cov}(\mathbf{Z})$,

$$\tilde{T}' \Pi \tilde{T} = \operatorname{diag}(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_k) = \tilde{D},$$

$\Pi\tilde{\mathbf{t}}_j = \lambda_j\tilde{\mathbf{t}}_j$ y $\mathbf{Z} = (Z_1, \ldots, Z_k)'$.

De esta forma, se obtiene que

$$\mathrm{Cov}(\tilde{\mathbf{Y}}) = \mathrm{Cov}(\tilde{T}'\mathbf{Z}) = \tilde{T}'\Pi\tilde{T} = \tilde{D}.$$

Es decir, las componentes principales obtenidas a partir de la matriz de correlaciones serán las variables incorreladas con varianza máxima que se pueden obtener a partir de combinaciones lineales de las variables estandarizadas

$$\mathbf{Z} = \mathrm{diag}(V)^{-\frac{1}{2}}(\mathbf{X} - \mu)$$

Sin embargo, los resultados que se obtienen son (en general) diferentes de los que se obtienen a partir de V.

- Proposición

Si $\tilde{\mathbf{Y}}$ son las componentes principales obtenidas a partir de la matriz de correlaciones de $\mathbf{X}$ entonces

$$\mathrm{Corr}(\mathbf{X}, \tilde{\mathbf{Y}}) = \tilde{T}\tilde{D}^{-\frac{1}{2}}.$$

- Demostración

En efecto, si $\tilde{\mathbf{Y}} = \tilde{T}'\mathbf{Z} = \tilde{T}'\mathrm{diag}(V)^{-\frac{1}{2}}(\mathbf{X} - \mu)$, entonces

$$\mathrm{Cov}(\mathbf{Z}, \tilde{\mathbf{Y}}) = \mathrm{Cov}(\mathbf{Z}, \tilde{T}'\mathbf{Z}) = \Pi\tilde{T} = \tilde{T}\tilde{D}.$$

$\mathrm{Corr}(\mathbf{X}, \tilde{\mathbf{Y}}) = \mathrm{Corr}(\mathbf{Z}, \tilde{\mathbf{Y}}) = \mathrm{Cov}(\mathbf{Z}, \tilde{\mathbf{Y}})\tilde{D}^{-\frac{1}{2}} = \tilde{T}\tilde{D}\tilde{D}^{-\frac{1}{2}} = \tilde{T}\tilde{D}^{\frac{1}{2}}$.

- Observación

Nótese que las correlaciones con la componente $\tilde{Y}_j$ son proporcionales al vector propio $\tilde{\mathbf{t}}_j$ (columnas de T) con constante de proporcionalidad $\lambda_j^{\frac{1}{2}}$ ($\mathrm{Corr}(X_i, Y_j) = \tilde{t}_{i,j}\lambda_j^{\frac{1}{2}}$) y que

$$\sum_{i=1}^{k}\mathrm{Corr}^2(X_i, \tilde{Y}_j) = \sum_{i=1}^{k}\tilde{t}_{i,j}\tilde{\lambda}_j = \tilde{\lambda}_j.$$

De forma similar, se define la matriz de saturaciones $\tilde{A} = \mathrm{Corr}(\mathbf{X}, \tilde{\mathbf{Y}})$, que verifica

$$\tilde{A}\tilde{A}' = \tilde{T}\tilde{D}^{\frac{1}{2}}\tilde{D}^{\frac{1}{2}}\tilde{T}' = \mathrm{Cov}(\mathbf{Z}) = \mathrm{Corr}(\mathbf{X})$$

y

$$\tilde{A}'\tilde{A} = \tilde{D}^{\frac{1}{2}}\tilde{T}'\tilde{T}\tilde{D}^{\frac{1}{2}} = \tilde{D}.$$

Es decir, la matriz de saturación es una matriz que factoriza Π junto a su traspuesta de forma que las multiplicamos al revés nos da una matriz diagonal.

Si estudiamos k variables (numéricas) en una determinada población usando una muestra de n individuos, tendremos una tabla de datos de la forma siguiente:

Datos	X_1	$\cdots$	X_k
$\mathbf{O}_1'$	$X_{1,1}$	$\cdots$	$X_{1,k}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\mathbf{O}_n'$	$X_{n,1}$	$\cdots$	$X_{n,k}$

$\mathbf{O}_1, \ldots, \mathbf{O}_n$: muestra aleatoria simple (formada por n vectores aleatorios columna independientes e idénticamente distribuidos) del vector aleatorio k-dimensional $\mathbf{X} = (X_1, \ldots, X_k)'$.

- En muchas ocasiones, podremos suponer normal.
- Sin embargo, otras veces prescindiremos de estas hipótesis y únicamente analizaremos una tabla de datos, tratando de condensar la información contenida en la misma y de analizar (de forma descriptiva) las relaciones entre las variables y los individuos.

- En la práctica

Así, en la práctica, tendremos que la matriz de covarianzas V es desconocida, por lo que tendremos que estimarla.

Y, una vez estimada, procederemos al cálculo de las componentes principales.

De esta forma, las componentes principales (y los valores de la matriz T) dependerán de los valores muestrales y, por lo tanto serán vectores aleatorios (con individuos distintos, obtendremos componentes distintas).

Lo mismo les ocurrirá a los valores propios (serán estimaciones de los verdaderos valores propios).

4.5.1) Estimación de la matriz de covarianzas

Para estimar V podemos utilizar la matriz de cuasi-covarianzas muestrales S calculada como

$$\begin{aligned}
&\mathbf{O}_l = (X_{l,1}, \ldots, X_{l,k})' \\
&\overline{X}_j = \frac{1}{n}\sum_{l=1}^{n} X_{l,j} \\
&\overline{\mathbf{O}} = (\overline{X}_1, \ldots, \overline{X}_k)' = \frac{1}{N}\sum_{l=1}^{n} \mathbf{O}_l \\
&S = \frac{1}{n-1}\sum_{l=1}^{n} (\mathbf{O}_l - \overline{\mathbf{O}})(\mathbf{O}_l - \overline{\mathbf{O}})' = (S_{i,j}) \\
&S_{i,j} = \frac{1}{n-1}\sum_{l=1}^{n} (X_{l,i} - \overline{X}_i)(X_{l,j} - \overline{X}_j).
\end{aligned}$$

También podemos usar la matriz de covarianzas muestrales

$$\hat{V} = \frac{n-1}{n} S.$$

Ambas tendrán los mismo vectores propios, y si n es grande, casi los mismo valores propios.

4.5.2) Cálculo a partir de una muestra

Como no conocemos V, la aproximaremos mediante S o $\hat{V}$, las diagonalizaremos (calcularemos los ejes de sus elipsoides) y podremos calcular las componentes principales definidas como sigue.

4.5.3) Definiciones

Llamaremos componentes principales muestrales a las variables $\hat{\mathbf{Y}} = \hat{T}'\mathbf{X}$, donde $\hat{T}$ es la matriz ortogonal que diagonaliza $S(\hat{V})$ y llamaremos valores propios muestrales $\hat{\lambda}_j$ a los valores propios de $S(\hat{V})$.

- Los valores de $\hat{T}$ serán las cargas (loadings) o coeficientes muestrales.
- Llamaremos puntuaciones muestrales (scores) a los valores que obtendríamos para cada individuo en las componentes muestrales

$$P_{l,j} = Y_j(O_j) = \hat{\mathbf{t}}_j' O_l.$$

4.5.4) Cálculo a partir de la matriz de correlaciones

Si optamos por calcular las componentes principales a partir de la matriz de correlaciones, como también es desconocida, en su lugar se usará la matriz de correlaciones (de Pearson) muestrales

$$\begin{aligned}
R &= \mathrm{diag}(S)^{-\frac{1}{2}} S \mathrm{diag}(S)^{-\frac{1}{2}} = (R_{i,j}) \\
R_{i,j} &= S_{i,j}(S_{i,i}S_{j,j})^{-\frac{1}{2}} = \hat{V}_{i,j}(\hat{V}_{i,i}, \hat{V}_{j,j})^{-\frac{1}{2}}.
\end{aligned}$$

Esto equivaldrá a estandarizar las variables iniciales restándoles sus medias muestrales y dividiéndolas por sus cuasivarianzas (es decir, hacer que todas tengan la misma variabilidad).

- En este caso, las puntuaciones se calcularán como:

$$P_{l,j} = Y_j(O_l) = \hat{\mathbf{t}}_j' O_l^*$$

donde $\hat{\mathbf{t}}_j$ es el vector propio j-ésimo de R.

Los datos estandarizados se obtienen (estiman) como

$$O_l^* = \left(\frac{X_{l,1} - \overline{X}_1}{S_1}, \ldots, \frac{X_{l,k} - \overline{X}_k}{S_k} \right)$$

siendo $S_i = \sqrt{S_{i,i}}$ la cuasidesviación típica de la variable X_i.

La cuasidesviación típica S_j puede ser reemplazada por la desviación típica muestral $\hat{V}_j = \sqrt{\hat{V}_{j,j}}$.

4.5.5) Caso de muestras grandes

Si n es grande, $\hat{V}$ y S son prácticamente iguales.

Si $\mathbf{X}$ es normal,

- $\hat{V}$ es máximo verosímil
- S es insesgado para V
- $(n-1)S$ tiene una distribución (en el muestreo) Wischart `WK(n - 1, V)`.

A partir de este resultado, se puede obtener la distribución exacta de los estimadores de los valores propios, pero ésta es bastante complicada.

- Proposición

Si $\hat{\theta}$ es máximo verosímil para θ, entonces $g(\hat{\theta})$ es máximo verosímil para $g(\theta)$.

4.5.6) Consecuencia

Si usamos $\hat{V}$ y todos sus valores propios son distintos, se obtendrán estimadores máximo verosímiles para $t_{i,j}$ y λ_j.

4.5.7) Caso de normalidad

Si $\mathbf{X}$ es normal, puede probarse que asintóticamente,

- la distribución conjunta de los estimadores de los valores propios es normal multivariante,.
- la distribución conjunta de los estimadores de los valores $t_{i,j}$ también lo es.
- además, ambas son independientes entre sí.

4.5.8) Cálculo de las componentes principales maximizando la varianza muestral

El cálculo de las componentes principales muestrales se puede enfocar de otra forma.

Se busca la variable $\mathbf{a}'\mathbf{X}$ (combinación lineal de las originales) con $\mathbf{a}'\mathbf{a} = 1$ que aplicada a los individuos de la muestra nos de una variable con varianza (o cuasivarianza) muestral máxima.

La puntuación o contador (scores) del individuo j en esta nueva variable será $\mathbf{a}'\mathbf{O}_j$, su media muestral será

$$\frac{1}{n}\sum_{j=1}^{n} \mathbf{a}'\mathbf{O}_j = \mathbf{a}'\frac{1}{n}\sum_{j=1}^{n} \mathbf{O}_j = \mathbf{a}'\overline{\mathbf{O}}$$

y su cuasivarianza será

$$\frac{1}{n-1}\sum_{j=1}^{n} (\mathbf{a}'\mathbf{O}_j - \mathbf{a}'\overline{\mathbf{O}})^2 = \frac{1}{n-1}\sum_{j=1}^{n} \mathbf{a}'(\mathbf{O}_j - \overline{\mathbf{O}})(\mathbf{O}_j - \overline{\mathbf{O}})'\mathbf{a} = \mathbf{a}'S\mathbf{a} \qquad (*)$$

cuyo máximo se alcanza si $\mathbf{a}$ es un vector propio del mayor de los valores propios de S.

De forma análoga, se procederá para el cálculo de las restantes componentes principales muestrales.

Sí, por inducción, suponemos que los primeros $i-1$ vectores propios $\hat{\mathbf{t}}_j$ de S nos dan las variables incorreladas con mayor varianza y buscamos maximizar la varianza muestral de $\mathbf{a}'\mathbf{O}_j$ (es decir $\mathbf{a}'S\mathbf{a}$) para $\mathbf{a}'\mathbf{a}=1$ haciendo que la covarianza muestral

$$\frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{a}'\mathbf{O}_j-\mathbf{a}'\overline{\mathbf{O}})(\hat{\mathbf{t}}_j\mathbf{O}_j-\hat{\mathbf{t}}_j'\overline{\mathbf{O}})=\mathbf{a}'S\hat{\mathbf{t}}_j$$

sea cero para $j=1,\ldots,i-1$.

Escribiendo $\mathbf{a}$ en función de la base de vectores propios y procediendo como en el teorema principal se obtiene que el óptimo es

$$\mathbf{a}=\hat{\mathbf{t}}_i.$$

De esta forma, podemos representar a los individuos mediante sus puntuaciones en las dos o tres primeras componentes manteniendo de ellos la mayor información (variabilidad o dispersión) posible (aunque $\mathbf{O}_1,\ldots,\mathbf{O}_n$ no sea una muestra aleatoria simple).

4.5.9) Interpretaicón geométrica: cálculo minimizando las distancias cuadráticas

Geométricamente, el espacio formado por las m primeras componentes y que pasa por el punto $\mathbf{O}$ será el espacio de dimensión m que minimiza la suma de las distancias al cuadrado de los individuos a dicho espacio (regresión perpendicular).

De esta forma, el ACP será como realizar una regresión mínimo cuadrática usando las distancias mínimas (regresión ortogonal) en lugar de las distancias verticales de la regresión clásica (para predecir Y en función de $\mathbf{X}$).

RELACIÓN DE PROBLEMAS: ANÁLISIS DE COMPONENTES PRINCIPALES
ANÁLISIS ESTADÍSTICO MULTIVARIANTE
GRADO EN CIENCIA E INGENIERÍA DE DATOS

1. Calcular las componentes principales para una variable bidimensional con matriz de covarianzas

$$V = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}.$$

   ¿Qué información contiene cada componente? Calcular la matriz de saturaciones e interpretar sus valores.

2. Calcular las componentes principales para una variable bidimensional con matriz de correlaciones

$$\Pi = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}.$$

   ¿Qué condiciones debe verificar r? Calcular la información que contiene cada componente.

3. Calcular las componentes principales para una variable bidimensional con matriz de covarianzas

$$\begin{pmatrix} 10 & -3 \\ -3 & 2 \end{pmatrix}.$$

   Calcular la matriz de saturaciones e interpretar sus valores.

4. Calcular la primera componente principal para una variable tridimensional con media cero y matriz de correlaciones

$$\begin{pmatrix} 1 & 0.8 & 0.8 \\ 0.8 & 1 & 0.8 \\ 0.8 & 0.8 & 1 \end{pmatrix}.$$

5. Calcular las componentes principales para una variable tridimensional con media cero y matriz de covarianzas

$$\Sigma = \begin{pmatrix} \beta^2+\delta & \beta & \beta \\ \beta & 1+\delta & 1 \\ \beta & 1 & 1+\delta \end{pmatrix}.$$

   (Indicación: $\Sigma - \delta I = (\beta, 1, 1)'(\beta, 1, 1)$).

6. Demostrar que si las varianzas iniciales son iguales entonces las componentes principales que se obtienen con la matriz de covarianzas son iguales a las que se obtienen con la matriz de correlaciones.

7. Calcular las componentes principales de k variables con media cero, varianza uno y correlaciones iguales a r. ¿Qué condiciones debe verificar r? Calcular la información que contiene cada componente.

8. Demostrar que las componentes principales no son invariantes por cambio de escala.

1) Calcular las componentes principales para una variable bidimensional con matriz de covarianzas

$$V = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}.$$

¿Qué información contiene cada componente? Calcular la matriz de saturaciones e interpretar sus valores.

$$|V - \lambda I| = \begin{vmatrix} 1-\lambda & 0.8 \\ 0.8 & 1-\lambda \end{vmatrix} = (1-\lambda)^2 - 0.8^2 = \lambda^2 - 2\lambda + 1 - 0.64 = \lambda^2 - 2\lambda + 0.36$$

$$\lambda = \frac{2 \pm \sqrt{(-2)^2 - 4 \cdot 1 \cdot 0.36}}{2 \cdot 1} = \frac{2 \pm 1.6}{2} = 1 \pm 0.8$$

$$\boxed{\begin{array}{l} \lambda_1 = 1.8 \\ \lambda_2 = 0.2 \end{array}}$$

1ª Componente: $Vx = \lambda_1 x$

$$\begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 1.8 \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \longrightarrow \begin{pmatrix} x_1 + 0.8x_2 \\ 0.8x_1 + x_2 \end{pmatrix} = \begin{pmatrix} 1.8x_1 \\ 1.8x_2 \end{pmatrix} \longrightarrow \begin{pmatrix} -0.8x_1 + 0.8x_2 \\ 0.8x_1 - 0.8x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \longrightarrow x_1 = x_2 \longrightarrow v = \alpha(1,1)'$$

$$t_1 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)'$$

$$Y_1 = \frac{1}{\sqrt{2}}X_1 + \frac{1}{\sqrt{2}}X_2$$

2ª Componente: $Vx = \lambda_2 x$

$$\begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0.2 \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \longrightarrow \begin{pmatrix} x_1 + 0.8x_2 \\ 0.8x_1 + x_2 \end{pmatrix} = \begin{pmatrix} 0.2x_1 \\ 0.2x_2 \end{pmatrix} \longrightarrow \begin{pmatrix} 0.8x_1 + 0.8x_2 \\ 0.8x_1 + 0.8x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \longrightarrow x_1 = -x_2 \longrightarrow v = \alpha(1,-1)'$$

$$t_2 = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)'$$

$$Y_2 = \frac{1}{\sqrt{2}}X_1 - \frac{1}{\sqrt{2}}X_2$$

$A = \text{Corr}(X,Y) = (\text{Corr}(X_i, Y_j))_{i,j}$

$$A = \text{Corr}(X,Y) = \underbrace{\text{diag}(V)^{-\frac{1}{2}}}_{\text{Id.}} T D^{\frac{1}{2}} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \cdot \begin{pmatrix} \sqrt{1.8} & 0 \\ 0 & \sqrt{1.2} \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{0.8}}{\sqrt{2}} & \frac{\sqrt{0.2}}{\sqrt{2}} \\ \frac{\sqrt{0.8}}{\sqrt{2}} & -\frac{\sqrt{0.2}}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} 0.9486 & 0.3162 \\ 0.9486 & -0.3162 \end{pmatrix}$$

$$\boxed{\underline{\text{Nota:}}\ \text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}}}$$

$$\text{Corr}(X_1, Y_1) = \frac{\text{Cov}(X_1, Y_1)}{\sqrt{1} \cdot \sqrt{1.8}}$$

$$\text{Cov}(X_1, Y_1) = \text{Cov}\left(X_1, \frac{1}{\sqrt{2}}X_1 + \frac{1}{\sqrt{2}}X_2\right) = \frac{1}{\sqrt{2}}\text{Cov}(X_1, X_1) + \frac{1}{\sqrt{2}}\text{Cov}(X_1, X_2) = \frac{1}{\sqrt{2}} + \frac{0.8}{\sqrt{2}} = \frac{1.8}{\sqrt{2}}$$

$$\text{Cov}(X_1, Y_2) = \text{Cov}\left(X_1, \frac{1}{\sqrt{2}}X_1 - \frac{1}{\sqrt{2}}X_1\right) = \frac{1}{\sqrt{2}} \cdot 1 - \frac{1}{\sqrt{2}} \cdot 0.8 = \frac{0.2}{\sqrt{2}}$$

$$\text{Cov}(X_2, Y_1) = \text{Cov}\left(X_2, \frac{1}{\sqrt{2}}X_1 + \frac{1}{\sqrt{2}}X_2\right) = \frac{1.8}{\sqrt{2}}$$

$$\text{Cov}(X_2, Y_2) = \text{Cov}\left(X_2, \frac{1}{\sqrt{2}}X_1 - \frac{1}{\sqrt{2}}X_2\right) = \frac{0.8}{\sqrt{2}} - \frac{1}{\sqrt{2}} = -\frac{0.2}{\sqrt{2}}$$

$Y = T'X$

$$Y = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \cdot \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

$a_{ij} = \text{Corr}(X_i, Y_j)$	Y_1	Y_2
X_1	0.9486	0.3162
X_2	0.9486	-0.3162

$a_{ij}^2 = \text{Corr}^2(X_i, Y_j)$	Y_1	Y_2	
X_1	0.9	0.1	1
X_2	0.9	0.1	1
	1.8	0.2	

2) Calcular las componentes principales para una variable bidimensional con matriz de correlaciones

$$\Pi = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}.$$

¿Qué condiciones debe verificar r? Calcular la información que contiene cada componente.

a) Si $r = 0$, $Y = X$

b) Si $r \neq 0$, para que Π sea definida positiva tiene que ocurrir que $1 - r^2 > 0$, $r^2 < 1$, es decir, $-1 < r < 1$

$$|\Pi - \lambda I| = \begin{vmatrix} 1-\lambda & r \\ r & 1-\lambda \end{vmatrix} = (1-\lambda)^2 - r^2 = \lambda^2 - 2\lambda + 1 - r^2$$

$$\lambda = \frac{2 \pm \sqrt{(-2)^2 - 4 \cdot (1 - r^2)}}{2} = \frac{2 \pm \sqrt{4 - 4 + 4r^2}}{2} = \frac{2 \pm 2r}{2} = \begin{cases} \lambda_1 = \dfrac{2+2r}{2} = 1 + r \\ \lambda_2 = \dfrac{2-2r}{2} = 1 - r \end{cases}$$

$$\boxed{\begin{aligned} \lambda_1 &= 1 + r \\ \lambda_2 &= 1 - r \end{aligned}}$$

Vectores propios:

1) $-1 < r < 1$

$\Pi v = \lambda_1 v$

$$\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = (1+r) \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \longrightarrow \begin{pmatrix} x_1 + rx_2 \\ rx_1 + x_2 \end{pmatrix} = \begin{pmatrix} x_1 + rx_1 \\ x_2 + rx_2 \end{pmatrix} \longrightarrow \begin{pmatrix} -rx1 + rx_2 \\ rx_1 - rx_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \longrightarrow V = \alpha(1,1)'$$

$$t_1 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right)$$

$\Pi v = \lambda_2 v$

$$\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = (1-r) \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \longrightarrow \begin{pmatrix} x_1 + rx_2 \\ rx_1 + x_2 \end{pmatrix} = \begin{pmatrix} x_1 - rx_2 \\ x_2 - rx_2 \end{pmatrix} \longrightarrow \left.\begin{aligned} x_1 + rx_2 &= (1-r)x_1 \\ rx_1 + x_2 &= (1-r)x_2 \end{aligned}\right\} \longrightarrow \begin{aligned} rx_2 &= -rx_1 \\ rx_1 &= -rx_2 \end{aligned} \longrightarrow$$

$v = \alpha(1, -1)'$

$$t_2 = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}} \right)$$

$$Y = T'X = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \cdot \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} X_1 + \frac{1}{\sqrt{2}} X_2 \\ \frac{1}{\sqrt{2}} X_1 - \frac{1}{\sqrt{2}} X_2 \end{pmatrix}$$

$Y = (Y_1, Y_2)$

$\text{Var}(Y_1) = \lambda = 1 + r \quad \dfrac{\lambda_1}{\lambda_1 + \lambda_2} = \dfrac{(1+r)}{2} \cdot 100\%$

$\text{Var}(Y_2) = \lambda = 1 - r \quad \dfrac{\lambda_2}{\lambda_1 + \lambda_2} = \dfrac{(1-r)}{2} \cdot 100\%$

2) $-1 < r < 0$, $\lambda_1 = 1 - r$, $\lambda_2 = 1 + r$

$$Y_1 = \frac{1}{\sqrt{2}} X_1 - \frac{1}{\sqrt{2}} X_2 \quad \text{Var}(Y_1) = 1 - r \quad \frac{\lambda_1}{\lambda_1 + \lambda_2} \cdot 100\% = \frac{1-r}{2} 100\%$$

$$Y_2 = \frac{1}{\sqrt{2}} X_1 + \frac{1}{\sqrt{2}} X_2 \quad \text{Var}(Y_2) = 1 + r \quad \frac{\lambda_2}{\lambda_1 + \lambda_2} \cdot 100\% = \frac{1+r}{2} 100\%$$

c) Si $r = \pm 1$

$r = 1$

$\Pi = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ no definida positiva

$|\Pi - \lambda I| = 0$

$\begin{vmatrix} 1-\lambda & 1 \\ 1 & 1-\lambda \end{vmatrix} = 0 \longrightarrow (1-\lambda^2) = 0 \longrightarrow 1 - 2\lambda + \lambda^2 - 1 = 0 \longrightarrow \lambda(\lambda - 2) \begin{cases} \lambda = 0 \\ \lambda = 2 \end{cases}$

$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 2 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \longrightarrow \left.\begin{matrix} x_1 + x_2 = 2x_1 \\ x_1 + x_2 = 2x_2 \end{matrix}\right\} \longrightarrow x_1 = x_2$

$v = \alpha(1,1)'$

$t_1 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)'$

$Y_1 = \frac{1}{\sqrt{2}} X_1 + \frac{1}{\sqrt{2}} X_2$

Si $r = -1$

$\Pi = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$

$|\Pi - \lambda I| = 0$

$\begin{vmatrix} 1-\lambda & -1 \\ -1 & 1-\lambda \end{vmatrix} = 0 \longrightarrow (1-\lambda)^2 - 1 = 0 \longrightarrow \begin{cases} \lambda = 0 \\ \lambda = 2 \end{cases}$

$\Pi v = \lambda v$

$\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 2 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

$\left.\begin{matrix} x_1 - x_2 = 2x_1 \\ -x_1 + x_2 = 2x_2 \end{matrix}\right\} \longrightarrow x_1 = -x_2$

$v = \alpha(1,-1)'$

$t_1 = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)$

$Y_1 = \frac{1}{\sqrt{2}} X_1 - \frac{1}{\sqrt{2}} X_2$

3) Calcular las componentes principales para una variable bidimensional con matriz de covarianzas

$$\begin{pmatrix} 10 & -3 \\ -3 & 2 \end{pmatrix}.$$

Calcular la matriz de saturaciones e interpretar sus valores.

$|V - \lambda I| = \begin{vmatrix} 10-\lambda & -3 \\ -3 & 2-\lambda \end{vmatrix} = (10-\lambda)\cdot(2-\lambda) - (-3)^2 = 20 - 10\lambda - 2\lambda + \lambda^2 - 9 = \lambda^2 - 12\lambda - 11$

$\lambda = \frac{12 \pm \sqrt{(-12)^2 - 4 \cdot 1 \cdot 11}}{2 \cdot 1} = \frac{12 \pm 10}{2} = \begin{cases} \frac{12+10}{2} = 11 \\ \frac{12-10}{2} = 1 \end{cases}$

$\lambda_1 = 11$

$\lambda_2 = 1$

1º Componente: $Vx = \lambda_1 x$

$\begin{pmatrix} 10 & -3 \\ -3 & 2 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 11 \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \longrightarrow \begin{pmatrix} 10x_1 - 3x_2 \\ -3x_1 + 2x_2 \end{pmatrix} = \begin{pmatrix} 11x_1 \\ 11x_2 \end{pmatrix} \longrightarrow \begin{pmatrix} -x_1 - 3x_2 \\ -3x_1 - 9x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \longrightarrow x_1 = -3x_2 \longrightarrow v = \alpha(1,-3)'$

$t_1 = \left(\frac{1}{\sqrt{10}}, -\frac{3}{\sqrt{10}}\right)'$

2ª Componente: $Vx = \lambda_2 x$

$\begin{pmatrix} 10 & -3 \\ -3 & 2 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 1 \cdot \begin{pmatrix} x_1 \\ x_1 \end{pmatrix} \longrightarrow \begin{pmatrix} 10x_1 - 3x_2 \\ -3x_1 + 2x_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \longrightarrow \begin{pmatrix} 9x_1 - 3x_2 \\ -3x_1 + x_2 \end{pmatrix} \longrightarrow x_2 = 3x_1 \longrightarrow v = \alpha(3,1)'$

$t_2 = \left(\frac{3}{\sqrt{10}}, \frac{1}{\sqrt{10}}\right)'$

$$T = (t_1|t_2) = \begin{pmatrix} \frac{1}{\sqrt{10}} & \frac{3}{\sqrt{10}} \\ -\frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} \end{pmatrix}$$

$$A = \text{diag}(V)^{-\frac{1}{2}} T D^{-\frac{1}{2}} = \begin{pmatrix} 10 & 2 \end{pmatrix}^{-\frac{1}{2}} \cdot \begin{pmatrix} \frac{1}{\sqrt{10}} & \frac{3}{\sqrt{10}} \\ -\frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} \end{pmatrix} \cdot \begin{pmatrix} 11 & 0 \\ 0 & 1 \end{pmatrix}^{\frac{1}{2}} = \begin{pmatrix} -0.0764 & -0.7708 \end{pmatrix} \cdot \begin{pmatrix} 11 & 0 \\ 0 & 1 \end{pmatrix}^{\frac{1}{2}} = \begin{pmatrix} -0.2534 & -0.7708 \end{pmatrix}$$

4) Calcular la primera componente principal para una variable tridimensional con media cero y matriz de correlaciones

$$\begin{pmatrix} 1 & 0.8 & 0.8 \\ 0.8 & 1 & 0.8 \\ 0.8 & 0.8 & 1 \end{pmatrix}.$$

Veamos que condiciones tienen que verificar los autovalores

$$\begin{pmatrix} 1 & 0.8 & 0.8 \\ 0.8 & 1 & 0.8 \\ 0.8 & 0.8 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \lambda \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

$$\left.\begin{array}{l} x_1 + 0.8x_2 + 0.8x_3 = \lambda x_1 \\ 0.8x_1 + x_2 + 0.8x_3 = \lambda x_2 \\ 0.8x_1 + 0.8x_2 + x_3 = \lambda x_3 \end{array}\right\} \longrightarrow \begin{array}{l} (1-\lambda)x_1 + 0.8x_2 + 0.8x_3 = 0 \\ 0.8x_1 + (1-\lambda)x_2 + 0.8x_3 = 0 \\ 0.8x_1 + 0.8x_2 + (1-\lambda)x_3 = 0 \end{array}$$

Para $1 - \lambda = 0.8$, tendremos que $x_1 + x_2 + x_3 = 0$

Consideremos $\lambda = 0.2$, y los vectores ortogonales $u = (0, -1, 1)'$ y $v = (-2, 1, 1)'$ que generan el espacio $x_1 + x_2 + x_3 = 0$.

Partimos de $u = (0, -1, 1)'$ y buscamos otro que se ortogonal, $v = (x_1, x_2, x_3)$, $\begin{array}{l} -x_2 + x_3 = 0 \\ x_1 + x_2 + x_3 = 0 \end{array} \longrightarrow x_2 = x_3,\ x_1 = 2x_2 \longrightarrow v = (-2, 1, 1)'$

Para ver cual sería el tercer autovector buscamos $w = (x_1, x_2, x_3)'$ orotogonal a los dos anteriores

5) Calcular las componentes principales para una variable tridimensional con media cero y matriz de covarianzas

$$\Sigma = \begin{pmatrix} \beta^2 + \delta & \beta & \beta \\ \beta & 1 + \delta & 1 \\ \beta & 1 & 1 + \delta \end{pmatrix}.$$

(Indicación: $\Sigma - \delta I = (\beta, 1, 1)'(\beta, 1, 1)$).

$$\Sigma = \begin{pmatrix} \beta^2 + \delta & \beta & \beta \\ \beta & 1 + \delta & 1 \\ \beta & 1 & 1 + \delta \end{pmatrix}$$

$\Sigma - \delta I = (\beta, 1, 1)'(\beta, 1, 1)$

$$(\beta, 1, 1)'(\beta, 1, 1) = \begin{pmatrix} \beta \\ 1 \\ 1 \end{pmatrix} (\beta, 1, 1) = \begin{pmatrix} \beta^2 & \beta & \beta \\ \beta & 1 & 1 \\ \beta & 1 & 1 \end{pmatrix}$$

$\Sigma(\beta, 1, 1)' - \delta(\beta, 1, 1)' = (\beta, 1, 1)'(\beta, 1, 1)(\beta, 1, 1)'$

$u = (\beta, 1, 1)$

$\Sigma u - \delta u = u \cdot (\beta^2 + 2)$

$\Sigma u = \delta u + (\beta^2 + 2)u = (\delta + \beta^2 + 2)u$

Hemos encontrado un valor propio $\lambda = \beta^2 + \delta + 2$ y el vector propio $u = (\beta, 1, 1)'$

Consideremos el vector ortogonal a u, $v = (0, 1, -1)'$

$Sigmav = (\delta I + (\beta, 1, 1)'(\beta, 1, 1))v = \delta v$

Entonces, $v = (0, 1, -1)'$ sería el vector propio asociado al valor propio $\lambda = \delta$.

Ahora habrá que buscar un vector w que sea ortogonal a u y a v

Sea $w=(x_1,x_2,x_3)'$ tal que $\left.\begin{array}{l} x_2-x_3=0 \\ \beta x_1+x_2+x_3=0 \end{array}\right\} \longrightarrow \begin{array}{l} x_2=x_3 \\ \beta x_1+2x_2=0 \longrightarrow x_2=-\dfrac{\beta x_1}{2} \end{array}$

$w=\left(1,-\dfrac{\beta}{2},-\dfrac{\beta}{2}\right)'$

$\Sigma w=\left(\delta I+(\beta,1,1)'(\beta,1,1)\right)w=\delta w$

Entonces $w=\left(1,-\dfrac{\beta}{2},-\dfrac{\beta}{2}\right)$ será el vector propio asociado al valor propio $\lambda=\delta$.

Para que la matriz sea semidefinida positiva δ debe ser ≥ 0. Con esta condición se tendrá que $\beta^2+\delta+2>0$.

$\lambda_1=\beta^2+\delta+2>\lambda_2=\lambda_3=\delta$

$$\mathbf{t}_1=\frac{1}{\sqrt{\beta^2+2}}\begin{pmatrix}\beta\\1\\1\end{pmatrix},\ \mathbf{t}_2=\frac{1}{\sqrt{2}}\begin{pmatrix}0\\1\\-1\end{pmatrix};\ \mathbf{t}_3=\frac{1}{\sqrt{1+\frac{\beta^2}{2}}}\begin{pmatrix}1\\-\frac{\beta^2}{2}\\-\frac{\beta^2}{2}\end{pmatrix}$$

Componentes principales: $Y=T'X\longrightarrow\begin{cases} Y_1=\dfrac{1}{\sqrt{\beta^2+2}}\cdot(\beta X_1+X_2+x_3) \\ Y_2=\dfrac{1}{\sqrt{2}}\cdot(X_2-X_3) \\ Y_3=\dfrac{1}{\sqrt{1+\frac{\beta^2}{2}}}\cdot\left(X_1-\dfrac{\beta}{2}X_2-\dfrac{\beta}{2}X_3\right)\end{cases}$

Proporción de variablidad explicada por cada componente principal.

$$\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3}\cdot 100\%=\frac{\beta^2+\delta+2}{\beta^2+3\delta+2}\cdot 100\%$$
$$\frac{\lambda_2}{\lambda_1+\lambda_2+\lambda_3}\cdot 100\%=\frac{\lambda_3}{\lambda_1+\lambda_2+\lambda_3}\cdot 100\%=\frac{\delta}{\beta^2+3\delta+2}\cdot 100\%$$

a) Si $\delta>0$, tendremos 3 componentes principales (Σ definida positiva)

b) Si $\delta=0$, podemos considerar solamente 1 (Σ semidefinida positiva)

Componente principal (Y_1)

6) Demostrar que si las varianzas iniciales son iguales entonces las componentes principales que se obtienen con la matriz de covarianzas son iguales a las que se obtienen con la matriz de correlaciones.

$$V=\begin{pmatrix}\sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1k}\\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2k}\\ \vdots & & & \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_k^2\end{pmatrix}=\left\{\sigma_1^2=\cdots=\sigma_k^2=\sigma^2\right\}=\begin{pmatrix}\sigma^2 & \sigma_{12} & \cdots & \sigma_{1k}\\ \sigma_{21} & \sigma^2 & & \vdots\\ \vdots & & \ddots & \vdots\\ \sigma_{k1} & \cdots & \cdots & \sigma^2\end{pmatrix}=\sigma^2\begin{pmatrix}1 & \frac{\sigma_{12}}{\sigma^2} & \cdots & \frac{\sigma_{1k}}{\sigma^2}\\ & 1 & & \\ & & \ddots & \\ \frac{\sigma_{k_1}}{\sigma^2} & \cdots & \cdots & 1\end{pmatrix}=\sigma^2 R$$

$$R=\begin{pmatrix}1 & \frac{\sigma_{12}}{\sigma_1\sigma_2} & \cdots & \frac{\sigma_{1k}}{\sigma_1\sigma_k}\\ & 1 & & \\ & & \ddots & \\ \frac{\sigma_{k_1}}{\sigma_1\sigma_k} & \cdots & \cdots & 1\end{pmatrix}=\left\{\sigma_1^2=\cdots=\sigma_k^2=\sigma^2\right\}=\begin{pmatrix}1 & \frac{\sigma_{12}}{\sigma^2} & \cdots & \frac{\sigma_{1k}}{\sigma^2}\\ & 1 & & \\ & & \ddots & \\ \frac{\sigma_{k_1}}{\sigma^2} & \cdots & \cdots & 1\end{pmatrix}$$

7) Calcular las componentes principales de k variables con media cero, varianza uno y correlaciones iguales a r. ¿Qué condiciones debe verificar r? Calcular la información que contiene cada componente.

Para $k=2$, $\Pi=\begin{pmatrix}1 & r\\ r & 1\end{pmatrix}$

Valores propios: $1-r, 1+r$

Para $k=3$, $\Pi=\begin{pmatrix}1 & r & r\\ r & 1 & r\\ r & r & 1\end{pmatrix}$

Tenemos que encontrar un vector u tal que $\Pi u=\lambda u$.

Sea $u=(x_1,x_2,x_3)'$

$$\begin{pmatrix}1 & r & r\\ r & 1 & r\\ r & r & 1\end{pmatrix}\cdot\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\lambda\cdot\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}\longrightarrow\left.\begin{array}{r} x_1+rx_2+rx_3=\lambda x_1\\ rx_1+x_2+rx_3=\lambda x_2\\ rx_1+rx_2+x_3=\lambda x_3\end{array}\right\}\longrightarrow\left.\begin{array}{r}(1-\lambda)x_1+rx_2+rx_3=0\\ rx_1+(1-\lambda)x_2+rx_3=0\\ rx_1+rx_2+(1-\lambda)x_3=0\end{array}\right\}$$

Considerando $1-\lambda=r$, tendremos la ecuación $rx_1+rx_2+rx_3=0$.

Si $r=0$, $\Pi=I$ y las componentes principales $Y=X$

Si $r\neq 0$, $x_1+x_2+x_3=0$, consideramos los vectores $u=(0,1,-1)'$ y $v=(-2,1,1)'$, y el tercer vector propio $w=(1,1,1)'$

$\Pi w=\begin{pmatrix}1 & r & r\\ r & 1 & r\\ r & r & 1\end{pmatrix}\cdot\begin{pmatrix}1\\1\\1\end{pmatrix}=\begin{pmatrix}1+2r\\1+2r\\1+2r\end{pmatrix}=(1+2r)\cdot\begin{pmatrix}1\\1\\1\end{pmatrix}\longrightarrow$ Tendremos que $\lambda=1+2r$ es eñ autovalor propio asociado a $w=(1,1,1)'$

Para que Π sea semidefinida positiva se debe verificar que $1-r\geq 0$, $1+2r\geq 0$ esto es, $r\geq 1$, $r\geq -0.5$

Caso 1: $0<r\leq 1$, $\lambda_1=1+2r, \lambda_2=\lambda_3=1-r$

$$\begin{array}{ll} Y_1=\frac{1}{\sqrt{3}}(X_1+X_2+X_3) & \frac{\lambda_1}{3}100\%=\frac{1+2r}{3}\cdot 100\% \\ Y_2=\frac{1}{\sqrt{6}}(-2X_1+X_2+X_3) & \frac{\lambda_2}{3}100\%=\frac{1-r}{3}\cdot 100\% \\ Y_3=\frac{1}{\sqrt{2}}(X_1-X_2) & \frac{\lambda_3}{3}100\%=\frac{1-r}{3}\cdot 100\% \end{array}$$

- Si $r=1$, podemos considerar solamente la pirmera componente principal (Y_1)

Caso 2: $-0.5\leq r<0$, $\lambda_1=\lambda_2=\lambda_3=1+2r$

$$\begin{array}{ll} \lambda_1=\frac{1}{\sqrt{6}}(-2X_1+X_2+X_3) & \frac{1-r}{3}\cdot 100\% \\ \lambda_2=\frac{1}{\sqrt{2}}(X_1-X_2) & \frac{1-r}{3}\cdot 100\& \end{array}$$

8) Demostrar que las componentes principales no son invariantes por cambio de escala.

Tema 5: Análisis discriminante

5.1) Introducción

5.1.1) Objetivo

Cómo clasificar individuos entre varios grupos a partir de sus medidas en diversas variables aleatorias.

- Para ello construiremos funciones discriminantes que servirán para decidir en qué población incluimos a cada sujeto.

Esta técnica se puede aplicar a muy diferentes situaciones.

- Diagnosis de enfermedades.
- Clasificación de individuos de diferentes especies.
- Diagnosis de autoría en obras de arte.
- Clasificación de perfiles de clientes (por ejemplo en la concesión de créditos), etc.

Cuando no se conozcan las características de las poblaciones en las que se pueden clasificar los individuos, necesitaremos disponer de una muestra de las variables en estudio de individuos de cada grupo (al menos dos individuos por cada grupo) y de las medidas de los elementos a clasificar en esas variables.

5.1.2) Criterios

La clasificación se basará en la distancia de Mahalanobis del individuo a cada una de las poblaciones (sus medias).

La utilización de esta distancia es equivalente bajo normalidad a la utilización del criterio de máxima verosimilitud, que clasificará a un individuo en donde sus medidas sean más probables (verosímiles), es decir, donde la función de densidad sea mayor.

Este segundo criterio permitirá la extensión de dicha clasificación a más de dos poblaciones con diferentes matrices de covarianzas incluso sin la necesidad de la normalidad de las mismas.

5.1.3) Distancia de Mahalanobis

La distancia de Mahalanobis del vector $\mathbf{x}$ al vector μ basada en la matriz V se define como

$$d_V(\mathbf{x}, \mu) = \sqrt{(\mathbf{x}-\mu)' V^{-1} (\mathbf{x}-\mu)}.$$

Si V es la matriz identidad, obtenemos la distancia Euclídea.

Caso de normalidad: La función de densidad se expresa como

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

para $\mathbf{x} \in \mathbb{R}^k$, donde μ es el vector de medias y V es la matriz de covarianzas.

- Las circunferencias para la distancia de Mahalanobis con centro en μ coincidirán con las curvas de nivel de la función de densidad ($f(x) = \text{cte.}$).

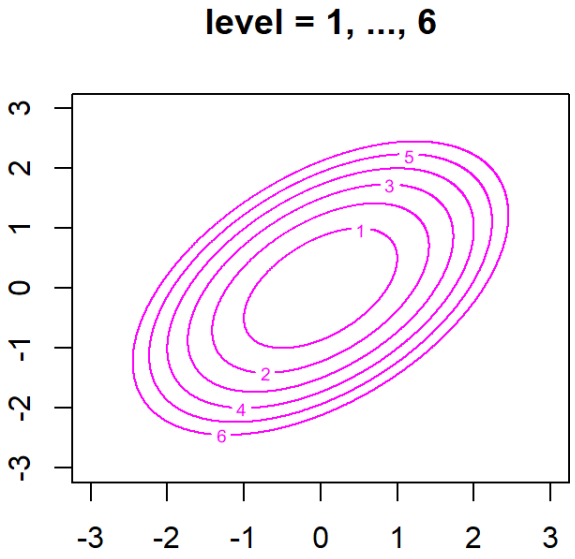
5.1.4) Para una normal bivariante

Por ejemplo, para una distribución

$$\mathcal{N}_2\left(\mu = \begin{pmatrix}0\\0\end{pmatrix}, V = \begin{pmatrix}1 & \frac{1}{2}\\ \frac{1}{2} & 1\end{pmatrix}\right)$$

Las circunferencias para la distancia de Mahalanobis con centro en μ coincidirán con las curvas de nivel de la función de densidad.

```
hc <- function(x1, x2) (4/3)*x1^ 2- (4/3)*x1*x2 + (4
    /3)*x2^ 2
x1 <- seq(-3, 3, length = 1000)
x2 <- seq(-3, 3, length = 1000)
z <- outer(x1, x2, hc)
contour(x1, x2, z, levels = c(1:6), col = "magenta")
title(main = "level = 1, ..., 6")
```



5.2) Dos poblaciones normales con la misma matriz de covarianzas

5.2.1) Clasificación teórica

Supongamos que $\mathbf{X} = (X_1, \ldots, X_k)'$ e $\mathbf{Y} = (Y_1, \ldots, Y_k)'$ son dos vectores aleatorios normales k-dimensionales con vectores de medias μ_X y μ_Y y matriz de covarianzas común V definida positiva.

Supongamos que $\mathbf{Z} = (Z_1, \ldots, Z_k)'$ representa las medidas obtenidas para el individuo que se quiere clasificar y que $\mathbf{Z}$ proviene de X o de $\mathbf{Y}$, es decir, $\mathbf{Z}$ será un vector aleatorio k-dimensional con media igual a μ_X o μ_Y y matriz de covarianzas V.

En la práctica $\mathbf{z}$ será un punto de $\mathbb{R}^k$ que debemos clasificar en $\mathbf{X}$ o en $\mathbf{Y}$.

La idea de Fisher es usar una función discriminante D (Función discriminante de Fisher) unidimensional lineal basada en $\mathbf{Z}$:

$$D = \mathbf{a}'\mathbf{Z} = a_1 Z_1 + \cdots + a_k Z_k,$$

donde $\mathbf{a} \in \mathbb{R}^k$.

Si $\mathbf{Z} \longrightarrow \mathcal{N}_k(\mu, V)$, entonces

$$D = \mathbf{a}'\mathbf{Z} \longrightarrow \mathcal{N}_1(\mathbf{a}'\mu,\ \mathbf{a}'V\mathbf{a})$$

ya que $E[\mathbf{a}'\mathbf{Z}] = \mathbf{a}'E[\mathbf{Z}]$ y

$$\mathrm{Var}(\mathbf{a}'\mathbf{Z}) = \mathrm{Cov}(\mathbf{a}'\mathbf{Z}) = \mathbf{a}'\,\mathrm{Cov}(\mathbf{Z})\mathbf{a} = \mathbf{a}'V\mathbf{a},$$

donde $\mu = E(\mathbf{Z}) = \mu_X$ o μ_Y.

Esta función debe elegirse de forma que discrimine (aleje) a los individuos de $\mathbf{X}$ de los de $\mathbf{Y}$.

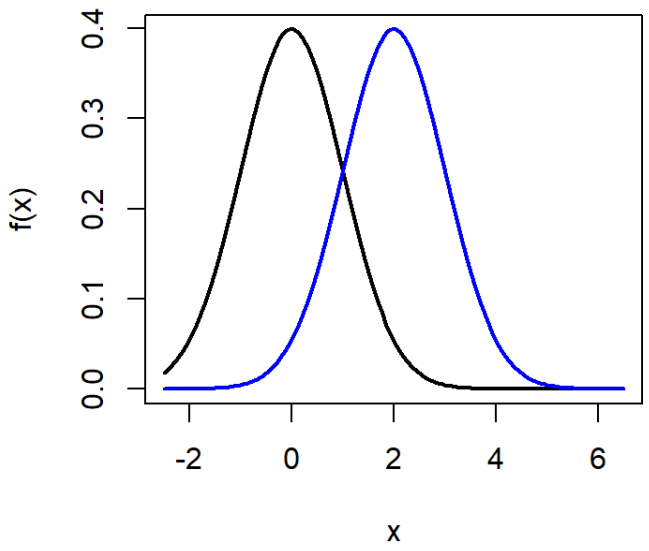
- Debemos resolver el problema siguiente:

$$\max_{a} \frac{(\mathbf{a}'\mu_X - \mathbf{a}'\mu_Y)^2}{\mathbf{a}'V\mathbf{a}}$$

- El objeto es alejar las proyecciones de las medias $\mathbf{a}'\mu_X$ y $\mathbf{a}'\mu_Y$ y disminuir la varianza común $\sigma^2 = \mathbf{a}'V\mathbf{a}$.

Ejemplo: funciones de densidad de las proyecciones en cada grupo.

```
x = seq(-2.5, 6.5, length.out = 100)
densidad_1 <- dnorm(x, mean = 0, sd = 1)
densidad_2 <- dnorm(x, mean = 2, sd = 1)
plot(x, densidad_1, type = "l", lwd = 2, col = "black",
     xlab = "x", ylab = "f(x)")
lines(x, densidad_2, type = "l", lwd = 2, col = "blue",
     xlab = "x", ylab = "f(x)", add = TRUE)
```



- Teorema

Si V es definida positiva, la solución general del problema

$$\max_{\mathbf{a}} \frac{(\mathbf{a}'\mu_X - \mathbf{a}'\mu_Y)^2}{\mathbf{a}'V\mathbf{a}}$$

viene dada por

$$\mathbf{a} = \lambda V^{-1}(\mu_X - \mu_Y)$$

para $\lambda \neq 0$, y el máximo vale $d_V^2(\mu_X, \mu_Y)$.

- Demostración

La demostración se basa en la desigualdad de Cauchy-Schwarz:

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

donde se da la igualdad si, y solo si, $\mathbf{x} = \lambda\mathbf{y}$.

Como V es definida positiva, existe su inversa V^{-1} y $\mathbf{a}'V\mathbf{a} > 0$ para todo vector $\mathbf{a} \neq 0$.

Entonces, tenemos

$$\begin{aligned}\frac{(\mathbf{a}'\mu_{\mathbf{X}} - \mathbf{a}'\mu_{\mathbf{Y}})^2}{\mathbf{a}'V\mathbf{a}} &= \frac{\left(\mathbf{a}'V^{\frac{1}{2}}(\mu_X - \mu_Y)\right)^2}{\mathbf{a}'V\mathbf{a}} \\ &\leq \frac{\mathbf{a}'V\mathbf{a}(\mu_X - \mu_Y)'V^{-1}(\mu_X - \mu_Y)}{\mathbf{a}'V\mathbf{a}} \\ &= (\mu_X - \mu_Y)'V^{-1}(\mu_X - \mu_Y) \\ &= d_V^2(\mu_X, \mu_Y),\end{aligned}$$

donde hemos considerado $\mathbf{x}' = \mathbf{a}'V^{\frac{1}{2}}$ e $\mathbf{y} = V^{-\frac{1}{2}}(\mu_X - \mu_Y)$.

Además, se verifica la igualdad si, y solo si $\mathbf{x} = \lambda\mathbf{y}$, es decir, si

$$V^{\frac{1}{2}}\mathbf{a} = \lambda V^{-\frac{1}{2}}(\mu_X - \mu_Y),$$

lo que implica que $\mathbf{a} = \lambda V^{-1}(\mu_X - \mu_Y)$.

5.2.2) Función discriminante de Fisher

Llamaremos función discriminante de Fisher a la variable aleatoria

$$D = L(\mathbf{Z}) = \mathbf{a}'\mathbf{Z} = (\mu_X - \mu_Y)'V^{-1}\mathbf{Z}.$$

Si las variables $\mathbf{X}$ e $\mathbf{Y}$ son normales, entonces la nueva variable D será normal

$$D \longrightarrow \mathcal{N}_1\left((\mu_X - \mu_Y)'V^{-1}\mu, d_V^2(\mu_X, \mu_Y)\right),$$

donde $\mu = E(\mathbf{Z})$ es igual a μ_X ó μ_Y.

Hemos considerado $\lambda = 1$, pero esto no influye en la clasificación ya que podemos tomar cualquier otro λ no nulo.

- Por ejemplo, si tomamos
$$\lambda = \frac{1}{\|a\|}$$
obtenemos una proyección en la dirección de $\mathbf{a}$.

5.2.3) Regla de discriminación

Consideramos la función discriminante de Fisher y $K = L\left(\frac{\mu_X + \mu_Y}{2}\right)$.

La regla de discriminación será:

- Si $L(\mathbf{Z}) > K$, entonces $\mathbf{Z}$ es clasificado en $\mathbf{X}$.
- Si $L(\mathbf{Z}) < K$, entonces $\mathbf{Z}$ es clasificado en $\mathbf{Y}$.

En realidad clasificamos a un individuo con características $\mathbf{z}$ según $\mathbf{a}'\mathbf{z}$ esté más cerca de $\mathbf{a}'\mu_X$ o de $\mathbf{a}'\mu_Y$, ya que, como

$$(\mu_X - \mu_Y)'V^{-1}(\mu_X - \mu_Y) \geq 0,$$

entonces

$$\mathbf{a}'\mu_X = (\mu_X - \mu_Y)'V^{-1}\mu_X \geq (\mu_X - \mu_Y)'V^{-1}\mu_Y = \mathbf{a}'\mu_Y,$$

es decir, con esta función discriminante, la proyección de la media de $\mathbf{X}$ será siempre mayor que la proyección de la media de $\mathbf{Y}$.

Ocurrirá lo mismo si tomamos $\lambda > 0$ y lo contrario si tomamos $\lambda < 0$.

De esta forma, se crean dos regiones en el conjunto de posibles valores de $\mathbf{Z}$:

- La región de individuos que serán clasificados en $\mathbf{X}$:

$$R_X = \{\mathbf{z} \in \mathbb{R}^{\mathbf{k}} : \mathbf{L}(\mathbf{z}) > K\}$$

- La región de individuos que serán clasificados en $\mathbf{Y}$:

$$R_Y = \{\mathbf{z} \in \mathbb{R}^{\mathbf{k}} : \mathbf{L}(\mathbf{z}) < K\}$$

5.2.4) ¿Cómo de buena es la función discriminante de Fisher obtenida?

La función discriminante de Fisher será mejor cuanto más alejadas estén las medias $\mathbf{a}'\mu_{\mathbf{X}}$ y $\mathbf{a}'\mu_Y$, y cuanto más pequeña sea la varianza $\mathbf{a}'V\mathbf{a}$.

Así, el cociente

$$\frac{(\mathbf{a}'\mu_X - \mathbf{a}'\mu_Y)^2}{\mathbf{a}'V\mathbf{a}} = (\mu_X - \mu_Y)'V^{-1}(\mu_X - \mu_Y) = d_V^2(\mu_X, \mu_Y)$$

(que no depende de λ) puede servir para comparar una función de discriminación con otra.

Nota:
La discriminación será buena si las **medias poblacionales están alejadas** según la distancia de Mahalanobis asociada a V.

5.2.5) Otro criterio para medir la bondad de un criterio de clasificación

Podemos calcular las probabilidades de malas (buenas) clasificaciones.

Si llamamos error tipo 1, e_1, al que clasifica a un individuo de la población $\mathbf{X}$ en la población $\mathbf{Y}$, entonces

$$\begin{aligned}
\Pr(e_1) &= \Pr(\mathbf{Z} \in R_Y | \mathbf{Z} \equiv \mathbf{X}) = \Pr(L(\mathbf{X}) < K) \\
&= \Pr\left(\mathbf{a}'\mathbf{X} < \mathbf{a}'\frac{\mu_X + \mu_Y}{2}\right) \\
&= \Pr\left(\frac{\mathbf{a}'\mathbf{X} - \mathbf{a}'\mu_{\mathbf{X}}}{\sqrt{\mathbf{a}'V\mathbf{a}'}} < \frac{\mathbf{a}'(\mu_Y - \mu_X)}{2\sqrt{a'V\mathbf{a}}}\right) \\
&= \Pr\left(U < \frac{(\mu_X - \mu_Y)'V^{-1}(\mu_Y - \mu_X)}{2\sqrt{(\mu_X - \mu_Y)'V^{-1}(\mu_X - \mu_Y)}}\right) \\
&= \Pr\left(U < -\frac{1}{2}d_V(\mu_X, \mu_Y)\right),
\end{aligned}$$

donde $U \longrightarrow N_1(0, 1)$

De forma análoga, si llamamos error tipo 2, e_2, al que clasifica a un individuo de la población $\mathbf{Y}$ en la población $\mathbf{X}$, entonces puede comprobarse que

$$\Pr(e_2) = \Pr(\mathbf{Z} \in R_X | \mathbf{Z} \equiv \mathbf{Y}) = \Pr\left(U > \frac{1}{2}d_V(\mu_X, \mu_Y)\right) = \Pr(e_1)$$

Por lo tanto, las probabilidades de clasificaciones erróneas son iguales y solo dependen de la distancia de Mahalanobis entre las medias de las poblaciones.

Lógicamente las probabilidades de clasificaciones correctas vienen dadas por:

$$\begin{aligned}
\Pr(c_1) &= \Pr(\mathbf{Z} \in R_X | \mathbf{Z} \equiv \mathbf{X}) = 1 - \Pr(e_1), \\
\Pr(c_2) &= \Pr(\mathbf{Z} \in R_Y | \mathbf{Z} \equiv \mathbf{X}) = 1 - \Pr(e_2),
\end{aligned}$$

y también son iguales.

Un caso sencillo

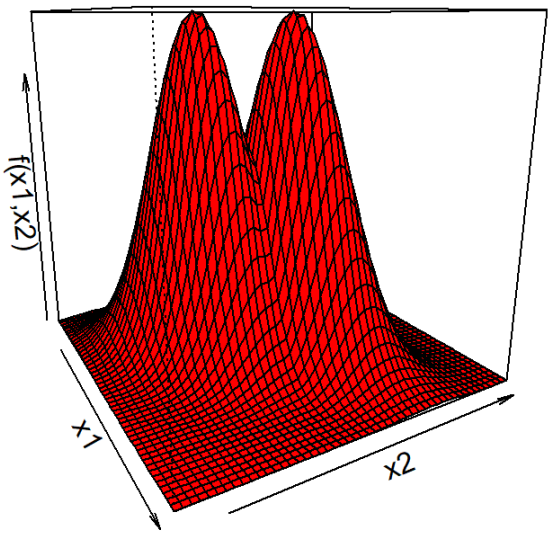
Supongamos que tenemos que decidir si un individuo con medidas

$$\mathbf{z} = (z_1, z_2)' = (2, 0.9)'$$

se clasifica en una población normal bivariante de media $\mu_X = (0, 0)'$ o en una de media $\mu_Y = (1, 2)'$ siendo la matriz de covarianzas común

$$V = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

```
library("mvtnorm")
V <- matrix(c(1, 1/2,
              1/2, 1), nrow = 2, ncol = 2, byrow = TRUE)
muX <- c(0, 0)
fX <- function(x1, x2) dmvnorm(data.frame(x1, x2), muX, V)
muY <- c(1, 2)
fY <- function(x1, x2) dmvnorm(data.frame(x1, x2), muY, V)
f <- function(x1, x2) pmax(fX(x1, x2), fY(x1, x2))
x <- seq(-3, 4, length = 50)
y <- seq(-3, 6, length = 50)
z <- outer(x, y, f)
persp(x, y, z, xlab = 'x1', ylab = 'x2', zlab = 'f(x1,x2)',
      col = 'red', theta = 60)
```



La función discriminante de Fisher será

$$D = L(\mathbf{Z}) = \mathbf{a}'\mathbf{Z} = (\mu_X - \mu_Y)'V^{-1}\mathbf{Z} = \begin{pmatrix} -1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}^{-1} \mathbf{Z}$$

$$= -\begin{pmatrix} 1 & 2 \end{pmatrix} \begin{pmatrix} \frac{4}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{4}{3} \end{pmatrix} \mathbf{Z} = -\begin{pmatrix} 0 & 2 \end{pmatrix} \mathbf{Z} = -2Z_2,$$

esto es, $L(z_1, z_2) = -2z_2$

La distancia de Mahalanobis al cuadrado entre las dos poblaciones vale

$$d_V^2(\mu_X, \mu_Y) = (\mu_X - \mu_Y)'V^{-1}(\mu_X - \mu_Y) = \begin{pmatrix} 0 & 2 \end{pmatrix} \begin{pmatrix} 1 & 2 \end{pmatrix}' = 4$$

Un individuo $\mathbf{Z}$ será clasificado en la primera población si

$$-2Z_2 > K = a' \frac{\mu_X + \mu_Y}{2} = -\begin{pmatrix} 0 & 2 \end{pmatrix} \begin{pmatrix} 0.5 & 1 \end{pmatrix}' = -2,$$

es decir, si $Z_2 < 1$

En este caso, $\mathbf{z} = \begin{pmatrix} 1 & 0.9 \end{pmatrix}$ será clasificado en $\mathbf{X}$, con una probabilidad de error global

$$\begin{aligned} \Pr(e_2) &= \Pr(\mathbf{Z} \in R_X | \mathbf{Z} \equiv \mathbf{Y}) \\ &= \Pr\left(U > \frac{1}{2} d_V(\mu_X, \mu_Y)\right) \\ &= \Pr(U > 1) \\ &= 1 - F_U(1) \\ &= 1 - 0.8413 = 0.1587 \end{aligned}$$

donde la función de distribución normal estándar $F_U(1)$ se puede calcular con las tablas estadísticas o en R con la instrucción `pnorm(1)`.

Otra función discriminante equivalente será

$$L^*(z_1, z_2) = z_2,$$

(proyección sobre el eje y) con la que obtendríamos

$$\begin{aligned} L^*(\mu_X) &= L^*(0,0) = 0, \\ L^*(\mu_Y) &= L^*(1,2) = 2, \\ K^* &= \frac{L^*(\mu_X) + L^*(\mu_Y)}{2} = 1 \end{aligned}$$

y

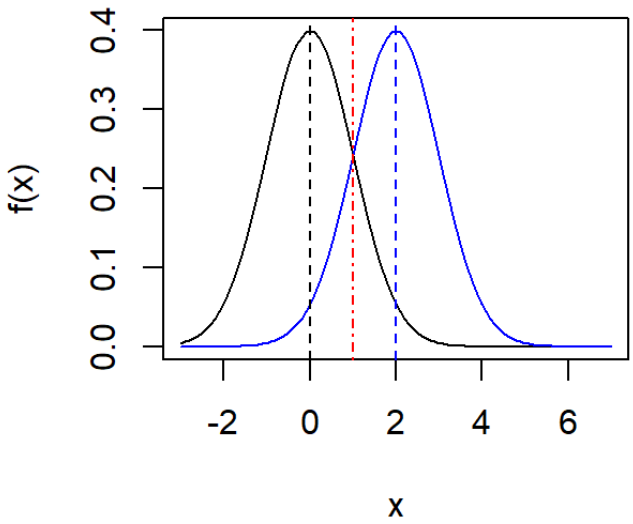
$$L^*(\mathbf{z}) = L^*(1, 0.9) = 0.9,$$

con lo que $\mathbf{z}$ se clasificará en $\mathbf{X}$.

Las proyecciones con esta función serán $N(0,1)(L^*(\mathbf{X}))$ y $N(2,1)(L^*(\mathbf{Y}))$.

Funciones de densidad de las proyecciones sobre el eje y en cada grupo para las poblaciones:

```
curve(dnorm(x, 0, 1), -3, 7, ylab = 'f(x)')
curve(dnorm(x, 2, 1), add = TRUE, col ='blue')
abline(v = 0, col = "black", lty = 2)
abline(v = 2, col = "blue", lty = 2)
abline(v = 1, col = "red", lty = 4)
```



La probabilidad de error 0.1587 corresponde a las áreas menores determinadas por la recta vertical en el punto de corte de las densidades en $K = 1$.

Nota:
Welch (1939) probó que, si las **poblaciones son normales**, el procedimiento de clasificación mediante la función discriminante de Fisher es máximo verosímil, es decir, se clasifica a un individuo con características $\mathbf{z}$ en $\mathbf{X}$ si y solo si $\mathbf{f_X}(\mathbf{z}) > \mathbf{f_Y}(\mathbf{z})$.
Además, se comprueba que también es equivalente al criterio de clasificación basado en la distancia de Mahalanobis mínima.

- Teorema

Si las variables $\mathbf{X}$ e $\mathbf{Y}$ son normales multivariantes con matriz de covarianzas común V y función discriminante de Fisher L, entonces equivalen:

1) $L(\mathbf{z}) > K$
2) $d_V(\mathbf{z}, \mu_\mathbf{X}) < d_V(\mathbf{z}, \mu_\mathbf{Y})$
3) $f_X(\mathbf{z}) > f_Y(\mathbf{z})$

- Demostración

La primera condición $L(\mathbf{z}) > K$ es

$$\mathbf{a}'\mathbf{z} > \frac{\mu_X + \mu_Y}{2},$$

con $\mathbf{a}' = (\mu_X - \mu_Y)'V^{-1}$, es decir,

$$\begin{aligned}(\mu_X - \mu_Y)'V^{-1}\mathbf{z} &> \frac{1}{2}(\mu_X - \mu_Y)'V^{-1}(\mu_X - \mu_Y)\\ &= \frac{1}{2}\mu_X'V^{-1}\mu_X - \frac{1}{2}\mu_Y'V^{-1}\mu_Y\end{aligned}$$

lo que equivale a

$$2\mu_X'V^{-1}\mathbf{z} - 2\mu_Y'V^{-1}\mathbf{z} > \mu_X'V^{-1}\mu_X - \mu_Y'V^{-1}\mu_Y.$$

Si las poblaciones son normales, la tercera opción es

$$c \cdot \exp\left(-\frac{1}{2}(\mathbf{z} - \mu_X)'V^{-}(\mathbf{z} - \mu_X)\right) > c \cdot \exp\left(-\frac{1}{2}(\mathbf{z} - \mu_Y)'V^{-}(\mathbf{z} - \mu_Y)\right),$$

donde $c = \dfrac{1}{\sqrt{|V|(2\pi)^k}}$. Es decir,

$$(\mathbf{z} - \mu_X)'V^{-}(\mathbf{z} - \mu_X) < (\mathbf{z} - \mu_Y)'V^{-}(\mathbf{z} - \mu_Y)$$

lo que es equivalente a la condición segunda.

- Observación

En la demostración anterior la hipótesis de normalidad no es necesaria para demostrar la equivalencia entre las dos primeras condiciones.

5.2.6) Diferente importancia a los dos tipos de errores

En ocasiones no es conveniente dar la misma importancia a los dos tipos de errores.

Podemos usar el criterio utilizado en los contrastes de hipótesis (Neyman-Pearson):

- Fijar un máximo para uno de los errores $\Pr(e_1) \leq \alpha$.
- Intentar reducir la probabilidad del otro error ($\Pr(e_2)$).

Usando este criterio sobre la función discriminante de Fisher, cambiará la constante K, que ahora se calculará a partir de la relación

$$\Pr(e_1) = \Pr(\mathbf{Z} \in \mathbf{R_Y} | \mathbf{Z} \equiv \mathbf{X}) = \Pr(L(\mathbf{X}) < K_\alpha) = \alpha,$$

donde $L(\mathbf{X}) \longrightarrow N_1\left((\mu_X - \mu_Y)'V^{-1}\mu_X, \sigma^2\right)$ y

$$\sigma^2 = d_V^2(\mu_X, \mu_Y) = (\mu_X - \mu_Y)'V^{-1}(\mu_X - \mu_Y).$$

De esta forma, la probabilidad del otro error valdrá

$$\Pr(e_2) = \Pr(L(\mathbf{Y}) > K_\alpha),$$

donde $L(\mathbf{Y}) \longrightarrow N_1\left((\mu_X - \mu_Y)' V^{-1} \mu_Y,\, \sigma^2\right)$.

5.2.7) Criterio de mínimo coste (probabilidad de error)

Otro criterio consiste en asignar un coste a cada uno de los posibles errores ($c_1, c_2 > 0$).

Supongamos que se concen las probabilidades a priori de pertenencia a cada una de las poblaciones:

- $q_1 = \Pr(\mathbf{Z} \equiv \mathbf{X})$
- $q_2 = \Pr(\mathbf{Z} \equiv \mathbf{Y})$

Entonces usando el teorema de la probabilidad total, se tiene

$$\begin{aligned}\Pr(\text{error}) =& \Pr(\mathbf{Z} \in R_Y | \mathbf{Z} \equiv \mathbf{X}) \Pr(\mathbf{Z} \equiv \mathbf{X}) \\ &+ \Pr(\mathbf{Z} \in R_X | \mathbf{Z} \equiv \mathbf{Y}) \Pr(\mathbf{Z} \equiv \mathbf{Y}) \\ =& \Pr(e_1) q_1 + \Pr(e_2) q_2\end{aligned}$$

Y el coste esperado asociado para una constante k será

$$c(k) = c_1 \Pr(e_1) q_1 + c_2 \Pr(e_2) q_2,$$

donde

$$\begin{aligned}\Pr(e_1) &= \Pr\left(L(\mathbf{X}) < k\right) = G\left(\frac{k - L(\mu_X)}{\sigma}\right), \\ \Pr(e_2) &= \Pr\left(L(\mathbf{Y}) < k\right) = 1 - G\left(\frac{k - L(\mu_Y)}{\sigma}\right),\end{aligned}$$

donde G es la función de distribución de la normal estándar $\mathcal{N}(0, 1)$.

Para minimizar el coste esperado, debe tomarse

$$k = \mathbf{a}' \frac{\mu_X + \mu_Y}{2} + \log\left(\frac{c_2 q_2}{c_1 q_1}\right).$$

5.2.8) Criterio de máxima probabilidad a posteriori

Supongamos que se conocen las probabilidades a priori q_1 y q_2.

Entonces se pueden calcular las probabilidades a posteriori (es decir, cuando conocemos los valores de $\mathbf{Z}$) para un individuo con medidas $\mathbf{z}$ mediante el Teorema de Bayes como:

$$\begin{aligned}\Pr(\mathbf{Z} \equiv \mathbf{X} | \mathbf{Z} = \mathbf{z}) &= \frac{\Pr(\mathbf{Z} = \mathbf{z} | \mathbf{Z} \equiv \mathbf{X}) \Pr(\mathbf{Z} \equiv \mathbf{X})}{\Pr(\mathbf{Z} = \mathbf{z})}, \\ \Pr(\mathbf{Z} \equiv \mathbf{Y} | \mathbf{Z} = \mathbf{z}) &= \frac{\Pr(\mathbf{Z} = \mathbf{z} | \mathbf{Z} \equiv \mathbf{Y}) \Pr(\mathbf{Z} \equiv \mathbf{Y})}{\Pr(\mathbf{Z} = \mathbf{z})},\end{aligned}$$

con

$$\Pr(\mathbf{Z} = \mathbf{z}) = \Pr(\mathbf{Z} = \mathbf{z} | \mathbf{Z} \equiv \mathbf{X}) \Pr(\mathbf{Z} \equiv \mathbf{X}) + \Pr(\mathbf{Z} = \mathbf{z} | \mathbf{Z} \equiv \mathbf{Y}) \Pr(\mathbf{Z} \equiv \mathbf{Y}).$$

Según este criterio, se clasificaría al individuo $\mathbf{z}$ en la población en la que tenga mayor probabilidad a posteriori.

5.3) Varias poblaciones con la misma matriz de covarianza

Cuando tengamos más de dos poblaciones con matriz de covarianzas común V, podemos usar el criterio de mínima distancia de Mahalanobis a las medias de los grupos.

$\mathbf{X^{(i)}}$ representan las diferentes poblaciones con medias $\mu^{(i)} = E(\mathbf{X}^{(i)})$ y la matriz de covarianzas V, para $i = 1, \ldots, m$.

Para un individuo con características $\mathbf{z}$ calcularemos

$$\begin{aligned} d_V^2(\mathbf{z}, \mu^{(i)}) &= (\mathbf{z} - \mu^{(i)})' V^{-1} (\mathbf{z} - \mu^{(i)}) \\ &= \mathbf{z}' V^{-1} \mathbf{z} - 2(\mu^{(i)})' V^{-1} \mathbf{z} + (\mu^{(i)})' V^{-1} \mu^{(i)}. \end{aligned}$$

Como la parte cuadrática $\mathbf{z}' V^{-1} \mathbf{z}$ es común, podemos quedarnos solo con la parte lineal (en realidad, su opuesta) dada por

$$L_i(\mathbf{z}) = (\mu^{(i)})' V^{-1} \mathbf{z} - \frac{1}{2} (\mu^{(i)})' V^{-1} \mu^{(i)}$$

conocida como función discriminante lineal (FDL), clasificándose un individuo con características $\mathbf{z}$ en el grupo en el que tenga un valor máximo dicha función discriminante.

- Teorema

Si $\mathbf{X}^{(i)}$ tienen medias $\mu^{(i)} = E(\mathbf{X}^{(i)})$ y matriz de covarianzas común V para $i = 1, \ldots, m$, entonces equivalen:

1) $L_i(\mathbf{z}) \geq L_j(\mathbf{z})$ para todo j.
2) $d_V^2(\mathbf{z}, \mu^{(\mathbf{i})}) \leq d_V^2(\mathbf{z}, \mu^{(\mathbf{i})})$ para todo j

- Corolario

Si solo hay dos grupos, este criterio de clasificación es equivalente a usar la función discriminante de Fisher.

- Demostración

La demostración del corolario es inmediata ya que demostramos que el criterio de mínima distancia de Mahalanobis era equivalente a usar la función discriminante de Fisher.

En este caso también podemos aplicar de máxima verosimilitud clasificando a $\mathbf{Z}$ en el grupo para el que $f_i(\mathbf{z})$ sea máxima, donde f_i representa la densidad del grupo i.

- Teorema

Si $\mathbf{X}^{(j)} \sim N(\mu^{(i)}, V)$ para $j = 1, \ldots, m$, entonces equivalen:

1) $L_i(\mathbf{z}) \geq L_j(\mathbf{z})$ para todo j.
2) $d_V^2(\mathbf{z}, \mu^{(i)}) \leq d_V^2(\mathbf{z}, \mu^{(j)})$ para todo j.
3) $f_i(\mathbf{z}) \geq f_j(\mathbf{z})$ para todo j.

- Demostración

La demostración es inmediata ya que la densidad normal multivariante del grupo i vale

$$f_i(\mathbf{z}) = \frac{1}{\sqrt{|V|(2\pi)^k}} \exp\left(-\frac{1}{2} (\mathbf{z} - \mu^{(i)})' V^{-1} (\mathbf{z} - \mu^{(i)}) \right)$$

y será máxima cuando la distancia de Mahalanobis al cuadrado

$$d_V^2(\mathbf{z}, \mu^{(i)}) = (\mathbf{x} - \mu^{(i)})' V^{-1} (\mathbf{x} - \mu^{(i)})$$

sea mínima.

Esto no será, en general, cierto si las poblaciones no son normales o si tienen distintas matrices de covarianzas.

- Proposición

Si todas las poblaciones son normales con matriz de covarianzas común V, entonces los criterios de clasificación de máxima verosimilitud y de mínima distancia de Mahalanobis son equivalentes a aplicar el criterio de discriminación de Fisher paso a paso tomando las poblaciones de dos en dos.

De esta forma, podríamos estudiar primero si $\mathbf{Z}$ se clasifica en la población 1 o en la 2.

En el segundo paso discriminaríamos entre la 3 y la ganadora del primer paso y así, sucesivamente.

Sin embargo, este método no se puede aplicar en la práctica ya que al discriminar entre las poblaciones 1 y 2 solo se utilizarán los individuos de estas poblaciones para estimar V.

Ejemplo

Supongamos que tenemos que decidir si un individuo con medidas $\mathbf{z} = (x, y)' = (1,\ 0.9)'$ se clasifica:

- en una población normal bivariante de media $(0, 0)'$.
- en una media $(1, 2)'$
- en una media $\left(-\frac{1}{2}, 1\right)$.

siendo la matriz de covarianzas común

$$V = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}.$$

$$L_i(\mathbf{z}) = (\mu^{(i)})' V^{-1} \mathbf{Z} - \frac{1}{2}(\mu^{(i)})' V^{-1} \mu^{(i)}$$

$$V = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix} \longrightarrow V^{-1} = \begin{pmatrix} \frac{4}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{4}{3} \end{pmatrix}$$

$$L_1(\mathbf{z}) = \begin{pmatrix} 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{4}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{4}{3} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} - \frac{1}{2} \begin{pmatrix} 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{4}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{4}{3} \end{pmatrix} \begin{pmatrix} 0 \\ 0 \end{pmatrix} = 0$$

$$L_2(\mathbf{z}) = \begin{pmatrix} 1 & 2 \end{pmatrix} \begin{pmatrix} \frac{4}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{4}{3} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} - \frac{1}{2} \begin{pmatrix} 1 & 2 \end{pmatrix} \begin{pmatrix} \frac{4}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{4}{3} \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 2 \end{pmatrix} = 2y - 2$$

$$L_3(\mathbf{z}) = \begin{pmatrix} -\frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} \frac{4}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{4}{3} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} - \frac{1}{2} \begin{pmatrix} -\frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} \frac{4}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{4}{3} \end{pmatrix} \begin{pmatrix} -\frac{1}{2} \\ 1 \end{pmatrix} = \begin{pmatrix} -\frac{4}{3} & \frac{5}{3} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} -\frac{2}{3} & \frac{5}{6} \end{pmatrix} \begin{pmatrix} -\frac{1}{2} \\ 1 \end{pmatrix} = -\frac{4}{3}x + \frac{5}{3}y - \frac{7}{6}$$

$$L_1(\mathbf{z}) = L_2(\mathbf{z}) \longleftrightarrow 2y - 2 = 0 \longleftrightarrow y = 1$$

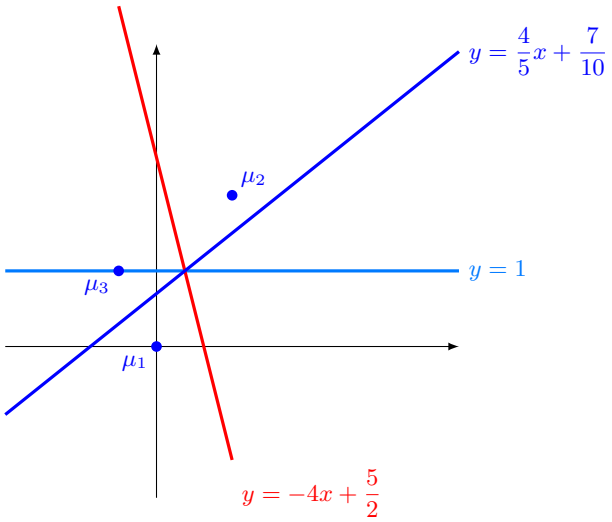
$$L_1(\mathbf{z}) = L_3(\mathbf{z}) \longleftrightarrow -\frac{4}{3}x + \frac{5}{3}y - \frac{7}{6} = 0 \longleftrightarrow y = \frac{3}{5}\left(\frac{4}{3}x + \frac{7}{6}\right) = \frac{4}{5}x + \frac{7}{10}$$

$$L_2(\mathbf{z}) = L_3(\mathbf{z}) \longleftrightarrow -\frac{4}{3}x + \frac{5}{3}y - \frac{7}{6} = 2y - 2 \longleftrightarrow 0 = 2y - 2 + \frac{4}{3}x - \frac{5}{3}y + \frac{7}{6} \longleftrightarrow \frac{4}{3}x + \frac{1}{3}y - \frac{5}{6} = 0$$

$$\longleftrightarrow y = \cancel{3} \cdot \left(-\frac{4}{\cancel{3}}x + \frac{5}{\cancel{6}2}\right) = -4x + \frac{5}{2}$$

$$\mathbf{z} = \begin{pmatrix} 1 & 0.9 \end{pmatrix}'$$

$$\left.\begin{array}{l} L_1(\mathbf{z}) = 0 \\ L_2(\mathbf{z}) = -0.2 \\ L_3(\mathbf{z}) = -1 \end{array}\right\} \text{Se clasifica en la población 1}$$



5.4) Varias poblaciones con distintas matrices de covarianza

5.4.1) Criterios de clasificación

Los criterios de clasificación por máxima verosimilitud o por mínima distancia de Mahalanobis a las medias de los grupos pueden utilizarse aunque las poblaciones no tengan la misma matriz de covarianzas.

No es necesario que las poblaciones sean normales, pudiéndose aplicar incluso a poblaciones de tipo discreto (siempre que se conozcan las densidades o las funciones puntuales de probabilidad).

Cuando las poblaciones sean normales suele hablarse de Análisis Discriminante Cuadrático (ADC o QDA) ya que las funciones que determinan las regiones de clasificación son polinomios de grado 2.

- Sin embargo, en este caso, las funciones discriminantes para mínima distancia o máxima verosimilitud no coinciden.

Bajo la hipótesis de normalidad, el criterio de máxima verosimilitud buscará el máximo de

$$f_i(\mathbf{z}) = \frac{1}{\sqrt{|V_i|(2\pi)^k}} \exp\left(-\frac{1}{2}(\mathbf{z}-\mu^{(i)})' V_i^{-1}(\mathbf{z}-\mu^{(i)})\right)$$

o, equivalentemente, el mínimo de

$$Q_i(\mathbf{z}) = c - 2\log f_i(\mathbf{z}) = (\mathbf{z}-\mu^{(i)})' V_i^{-1}(\mathbf{z}-\mu^{(i)}) + \log|V_i|,$$

con término constante $c = -l\log(2\pi)$, conocida como función discriminante cuadrática (QDF) para $i = 1, \ldots, m$.

- Un individuo con medidas $\mathbf{z}$ se clasificará en el grupo donde la función discriminante cuadrática sea mínima (máxima verosimilitud).

El criterio basado en la distancia de Mahalanobis usará la función discriminante cuadrática

$$Q_i^*(\mathbf{z}) = d^2_{V_i}(\mathbf{z}, \mu^{(\mathbf{i})}) = (\mathbf{z}-\mu^{(\mathbf{i})})' V_i^{-1}(\mathbf{z}-\mu^{(\mathbf{i})}).$$

5.4.2) Observaciones

Los resultados pueden ser diferentes (cuando los determinantes de las matrices de covarianzas sean diferentes).

En general, las funciones discriminantes cuadráticas son muy sensibles cuando las poblaciones no son normales, por lo que no es muy recomendable su uso en este caso, siendo preferible usar funciones discriminantes lineales.

Ejemplo

Sean dos poblaciones normales bidimensionales con medias $\mu_1 = (2,0)'$ y $\mu_2 = (0,0)'$, y matrices de covarianzas

$$V_1 = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}, \qquad \text{y} \qquad V_2 = \begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}.$$

Se pide:

- Calcular las funciones discriminantes cuadráticas.
- Proporcionar el criterio de clasificación.
- Dibujar las regiones de clasificación en $\mathbb{R}^2$.
- Clasificar a $\mathbf{z} = (1,1)'$

Como

$$V_1^{-1} = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix},$$

la primera QDF es

$$\begin{aligned} Q_1(x,y) &= \begin{pmatrix} x-2 & y \end{pmatrix} V_1^{-1} \begin{pmatrix} x-2 \\ y \end{pmatrix} \\ &= (x-2+y)(x-2) + (x-2+2y)y \\ &= x^2 - 5x + 4 + 2yx - 4y + 2y^2 \end{aligned}$$

Análogamente, para el segundo grupo tenemos:

$$\begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 2 \end{pmatrix}$$

con lo que la segunda QDF será

$$Q_2(x,y) = \begin{pmatrix} x & y \end{pmatrix} V_2^{-1} \begin{pmatrix} x \\ y \end{pmatrix} = 0.5x^2 + 2y^2$$

5.5) Clasificación a partir de una muestra

En la práctica, los valores de las medias y las matrices de covarianzas teóricos usados en los criterios de clasificación serán desconocidos, por lo que tendrán que ser estimados.

Estas estimaciones dependerán de las hipótesis de partida (normalidad, igualdad de matrices de covarianzas, etc...), hipótesis que en muchos casos deberán ser corroboradas mediante algún procedimiento cuando no se está muy seguro de su validez.

En general, si estudiamos k variables numéricas $(Z_1, \ldots, Z_k)$ en m poblaciones distintas indicadas por una variable discreta G (grupo) para una muestra de n individuos (muestra de entrenamiento), tendremos una tabla de datos de la forma

	Z_1	$\cdots$	Z_k	G
ω_1	$z_{1,1}$	$\cdots$	$z_{1,k}$	g_1
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
ω_1	$z_{1,1}$	$\cdots$	$z_{1,k}$	g_1

donde $g_i \in \{1, \ldots, m\}$ para todo i.

Para cada valor de $G = j$, esta tabla proporcionará una muestra aleatoria simple de la variable aleatoria k-dimensional $\mathbf{Z} = (Z_1, \ldots, Z_k)'$ condicionada por $G = j$ que, en muchas ocasiones, podremos suponer normal.

- En realidad tendremos m muestras de m poblaciones normales k-dimensionales $\mathcal{N}_k(\mu^{(i)}, V_j)$.

Así en la práctica, tendremos m medias teóricas y m matrices de covarianzas desconocidas, por lo que tendremos que estimarlas.

Para dichas estimaciones usaremos:

$$\hat{\mu}^{(j)} = \frac{1}{n_j}\sum_{i=1}^{n}\omega_i 1(G(\omega_i)=j)$$
$$\hat{V}_j = \frac{1}{n_j-1}\sum_{i=1}^{n}1(G(\omega_i)=j)(\omega_i-\hat{\mu}^{(j)})(\omega_i-\hat{\mu}^{(j)})'$$
$$\omega_i = (z_{i,1},\ldots,z_{i,k})'$$
$$n_j = \sum_{i=1}^{n}1(G(\omega_i)=j)$$

donde $n = n-1+\cdots+n_m$, y $1(G(\omega_i)=j)$ es la función indicador que vale 1 si el individuo i pertenece a la clase j-ésima y vale 0 en caso contrario. Y por lo tanto, n_j es el número de individuos de la muestra pertenecientes a la población j-ésima.

Necesitamos $n_j > 1$ para todo j.

Si $(\mathbf{Z}|G=j)$ es normal, $\hat{V}_j$ es insesgado para V_j, teniendo $(n_j-1)\hat{V}_j$ una distribución (en el muestreo) Wishart $W_k(n_j-1, V-j)$.

Si suponemos que las matrices de covarianzas teóricas son todas iguales ($V_1=\cdots=V_m=V$), entonces la matriz de covarianzas común V se aproximará mediante la matriz de covarianzas ponderada (`pooled`)

$$\hat{V} = \frac{1}{n-m}\sum_{j=1}^{m}(n_j-1)\hat{V}_j$$

que será un estimador insesgado para V.

A partir de estos estimadores se pueden obtener estimaciones de las distintas funciones discriminantes y con ellas, obtener clasificaciones (empíricas) para nuevos individuos.

Como los estimadores se aproximan a los verdaderos valores de los parámetros, las clasificaciones se parecerán (si n es grande) a las que se obtendrían usando los verdaderos parámetros.

Por ejemplo, para el caso de dos poblaciones con la misma matriz de covarianzas, la función discriminante de Fisher se estimará mediante

$$\hat{D} = \hat{L}(\mathbf{Z}) = \hat{a}'\mathbf{Z} = (\hat{\mu}_X-\hat{\mu}_Y)'\hat{V}^{-1}\mathbf{Z}.$$

- De esta forma, la probabilidad del error tipo 1 se estimará mediante

$$\Pr(e_1) = \Pr\left(U < -\frac{1}{2}\hat{d_V}\right),$$

donde $U \equiv N_1(0,1)$ y

$$\hat{d_V} = \sqrt{(\hat{\mu}_X-\hat{\mu}_Y)'\hat{V}^{-1}(\hat{\mu}_X-\hat{\mu}_Y)}$$

es la distancia de Mahalanobis muestral entre las medias (muestrales) de los grupos.

Bajo la hipótesis de normalidad, estos estimadores son asintóticamente insesgados.

En Srivastava y Carter (1983, pag. 238) pueden verse otros estimadores basados en las distribuciones obtenidas para los distintos errores a partir de la hipótesis de normalidad.

Se procederá de forma similar en los otros casos.

Por ejemplo, las Funciones Discriminantes Lineales (FDL) muestrales serán

$$\hat{L}_i(\mathbf{z}) = (\hat{\mu}^{(i)})'\hat{V}^{-1}\mathbf{z} - \frac{1}{2}\left((\hat{\mu}^{(i)})'\hat{V}^{-1}\hat{\mu}^{(i)}\right),$$

clasificándose $\mathbf{Z}$ en G_i si $\hat{L}_i(\mathbf{z}) \geq \hat{L}_j(\mathbf{z})$ para todo j.

Análogamente, las funciones discriminantes cuadráticas (FDC o QDF) muestrales serán

$$\hat{Q}_i(\mathbf{z}) = c - 2\log\hat{f}_i(\mathbf{z}) = (\mathbf{z}-\hat{\mu}^{(i)})'\hat{V}^{(-1)}(\mathbf{z}-\hat{\mu}^{(i)}) + \log|\hat{V}_i|,$$

clasificándose $\mathbf{Z}$ en G_i si $\hat{Q}_i(\mathbf{z}) \leq \hat{Q}_j(\mathbf{z})$ para todo j.

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

1. Dadas tres poblaciones normales bidimensionales con medias $\boldsymbol{\mu}^{(1)} = (1,0)'$, $\boldsymbol{\mu}^{(2)} = (0,1)'$ y $\boldsymbol{\mu}^{(3)} = (0,0)'$ y matrices de covarianzas iguales a

$$V = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix},$$

se pide:

   a) Obtener las funciones discriminantes lineales.
   b) Clasificar a $\mathbf{z} = (2,2)'$.
   c) Dibujar las regiones de clasificación para cada grupo.

2. Dadas tres poblaciones normales bidimensionales con medias $\boldsymbol{\mu}^{(1)} = (0,0)'$, $\boldsymbol{\mu}^{(2)} = (1,1)'$ y $\boldsymbol{\mu}^{(3)} = (2,0)'$ y matrices de covarianzas iguales a

$$V = \begin{pmatrix} 5 & 2 \\ 2 & 1 \end{pmatrix},$$

se pide:

   a) Obtener las funciones discriminantes lineales.
   b) Clasificar a $\mathbf{z} = (1,1/2)'$.
   c) Obtener la función discriminante de Fisher, la constante K y el criterio de clasificación para distinguir entre las poblaciones 2 y 3.
   d) Dibujar las regiones de clasificación para cada grupo.

3. Dadas tres poblaciones normales con matriz de covarianzas común

$$V = \begin{pmatrix} 6 & 2 \\ 2 & 1 \end{pmatrix}$$

y medias $(0,1), (1,0)$ y $(1,1)$, respectivamente, obtener las funciones discriminantes y el criterio de clasificación.

4. Dados dos vectores aleatorios normales bidimensionales con medias $(0,0)$ y $(3,0)$ y matrices de covarianzas

$$V_1 = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \text{ y } V_2 = \begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix},$$

se pide:

   a) Calcular las funciones discriminantes cuadráticas.

   b) Clasificar a $\mathbf{z} = (1,-4)'$ usando dichas funciones.

   c) Representar gráficamente las regiones de clasificación.

5. Dadas dos poblaciones normales bidimensionales con medias $\boldsymbol{\mu}^{(1)} = (1,0)'$ y $\boldsymbol{\mu}^{(2)} = (0,0)'$ y matrices de covarianzas

$$V_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1/2 \end{pmatrix} \text{ y } V_2 = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix},$$

se pide:

   a) Obtener las funciones discriminantes cuadráticas y clasificar a $\mathbf{z} = (1,1)'$.

   b) Clasificar a $\mathbf{z}$ usando el criterio de mínima distancia de Mahalanobis y representar las regiones de clasificación con este criterio para cada grupo.

6. Obtener un criterio de clasificación para dos poblaciones exponenciales unidimensionales con medias distintas usando máxima verosimilitud. Clasificar a $z = 1.5$ entre dos poblaciones exponenciales con medias 2 y 1. (Indicación: La función de densidad de la distribución exponencial es $f(x) = (1/\mu)\exp(-x/\mu)$ para $x \geq 0$).

1) Dadas tres poblaciones normales bidimensionales con medias $\mu^{(1)})(1,0)'$, $\mu^{(2)} = (0,1)'$ y $\mu^{(3)} = (0,0)'$ y matrices de covarianzas iguales a

$$V = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

se pide:

a) Obtener las funciones discriminantes lineales.

$$\mu^{(1)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$\mu^{(2)} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$\mu^{(3)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$V = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

$$|V| = 2 - 1 = 1$$

$$V^{-1} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$$

$$L_i(z) = \left(\mu^{(i)}\right)' V^{-1} z - \frac{1}{2}\left(\mu^{(i)}\right)' V^{-1} \mu^{(i)}$$

$$L_1(z) = \begin{pmatrix} 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} - \frac{1}{2}\begin{pmatrix} 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 & -1 \end{pmatrix}\cdot\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} - \frac{1}{2}\begin{pmatrix} 1 & -1 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = z_1 - z_2 - \frac{1}{2}$$

$$L_2(z) = \begin{pmatrix} 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} - \frac{1}{2}\begin{pmatrix} 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 & 2 \end{pmatrix}\cdot\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} - \frac{1}{2}\begin{pmatrix} -1 & 2 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \end{pmatrix} = -z_1 + 2z_2 - 1$$

$$L_3(z) = \begin{pmatrix} 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} - \frac{1}{2}\begin{pmatrix} 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} 0 \\ 0 \end{pmatrix} = 0$$

Criterio: $\mathbf{z}$ se clasificará en la población donde $L_i(\mathbf{z})$ sea máxima

b) Clasificar a $\mathbf{z} = (2,2)'$

$$\text{Para } \mathbf{z} = (2,2)' \longrightarrow \begin{cases} L_1(\mathbf{z}) = -\frac{1}{2} \\ L_2(\mathbf{z}) = 1 \\ L_3(\mathbf{z}) = 0 \end{cases}$$

Como $L(\mathbf{z}) = -2 < -\frac{1}{2}$, clasificamos el individuo $\mathbf{z}$ en la población Y.

c) Dibujar las regiones de clasificación para cada grupo

$$L_1(z) = L_2(z) \longrightarrow z_1 - z_2 - \frac{1}{2} = -z_1 + 2z_2 - 1$$

$$-z_2 - 2z_2 = -z_1 - z_1 - 1 + \frac{1}{2}$$

$$-3z_2 = -2z_1 - \frac{1}{2}$$

$$z_2 = \frac{2}{3}z_1 + \frac{1}{6}$$

$$L_2(z) = L_3(z) = 0$$

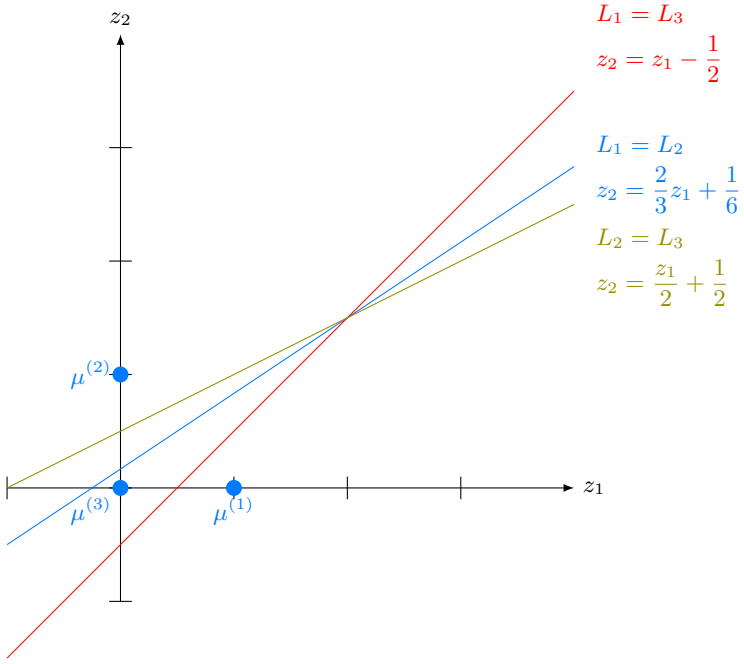
$$-z_1 + 2z_2 - 1 = 0$$

$$z_2 = \frac{z_1}{2} + \frac{1}{2}$$

$$L_1(z) = L_3(z) = 0$$

$$z_1 - z_2 - \frac{1}{2} = 0$$

$$z_2 = z_1 - \frac{1}{2}$$



2) Dadas tres poblaciones normales bidimensionales con medias $\mu^{(1)} = (0,0)'$, $\mu^{(2)} = (1,1)'$ y $\mu^{(3)} = (2,0)'$ y matrices de covarianzas iguales a

$$V = \begin{pmatrix} 5 & 2 \\ 2 & 1 \end{pmatrix},$$

se pide:

a) Obtener las funciones discriminantes lineales.

$$\mu^{(1)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\mu^{(2)} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\mu^{(3)} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$$

$$V = \begin{pmatrix} 5 & 2 \\ 2 & 1 \end{pmatrix}$$

$$|V| = 5 - 4 = 1$$

$$V^{-1} = \begin{pmatrix} 1 & -2 \\ -2 & 5 \end{pmatrix}$$

$$L_i(\mathbf{z}) = \left(\mu^{(i)}\right)' V^{-1}\mathbf{z} - \frac{1}{2}\left(\mu^{(i)}\right)' V^{-1}\mu^{(i)} \qquad \text{(FDL)}$$

$$L_1(\mathbf{z}) = \begin{pmatrix} 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & -2 \\ -2 & 5 \end{pmatrix}\mathbf{z} - \frac{1}{2}\begin{pmatrix} 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & -2 \\ -2 & 5 \end{pmatrix}\begin{pmatrix} 0 \\ 0 \end{pmatrix} = 0$$

$$L_2(\mathbf{z}) = \begin{pmatrix} 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & -2 \\ -2 & 5 \end{pmatrix}\mathbf{z} - \frac{1}{2}\begin{pmatrix} 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & -2 \\ -2 & 5 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = -z_1 + 3z_2 = 1$$

$$L_3(\mathbf{z}) = \begin{pmatrix} 2 & 0 \end{pmatrix}\begin{pmatrix} 1 & -2 \\ -2 & 5 \end{pmatrix}\mathbf{z} - \frac{1}{2}\begin{pmatrix} 2 & 0 \end{pmatrix}\begin{pmatrix} 1 & -2 \\ -2 & 5 \end{pmatrix}\begin{pmatrix} 2 \\ 0 \end{pmatrix} = 2z_1 - 4z_2 - 2$$

Criterio: z se clasificará en la población donde $L_i(\mathbf{z})$ sea máxima.

b) Clasificar a $\mathbf{z} = \left(1, \frac{1}{2}\right)'$.

$$\text{Para } z = \left(1, \frac{1}{2}\right)' \longrightarrow \begin{cases} L_1(\mathbf{z}) = 0 \\ L_2(\mathbf{z}) = -1 + \frac{3}{2} - 1 = -\frac{1}{2} \\ L_3(\mathbf{z}) = -2 - \frac{4}{2} - 2 = -2 \end{cases} \qquad \text{Se clasifica en la población (1)}$$

c) Obtener la función discriminante de Fisher, la constante K y el criterio de clasificación para distinguir entre las poblaciones 2 y 3.

Función discriminante de Fisher: $D = L(\mathbf{z}) = \mathbf{a}'\mathbf{Z} = (\mu^2 - \mu^{(3)})'V^{-1}\mathbf{z}$

$$L_(\mathbf{z}) = \begin{pmatrix} -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & -2 \\ -2 & 5 \end{pmatrix} \mathbf{z} = \begin{pmatrix} -3 & 7 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = -3z_1 + 7z_2$$

$$k = L\left(\frac{\mu^{(2)} + \mu^{(3)}}{2}\right) = L\left((1.5, 0.5)'\right) = -3 \cdot 1.5 + 7 \cdot 0.5 = -1$$

$$R_{\mu^{(2)}} = \{\mathbf{z} : L(\mathbf{z}) > k\} = \{z : -3z_1 + 7z_2 > -1\}$$

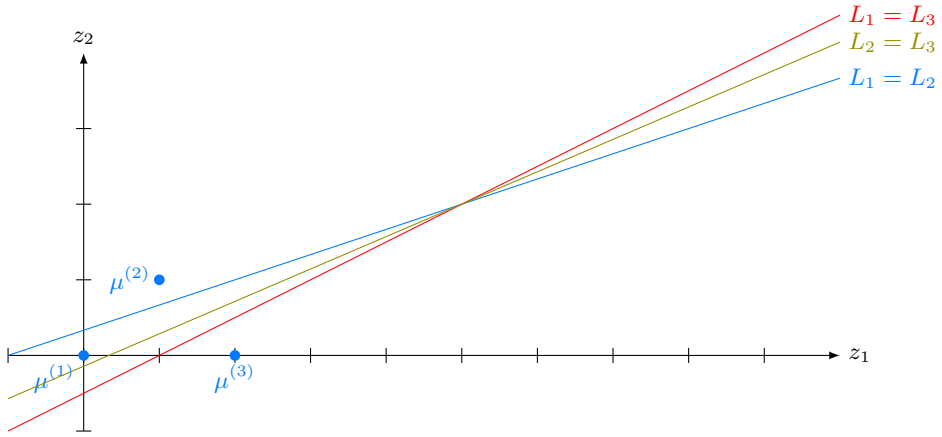
$$R_{\mu^{(3)}} = \{\mathbf{z} : L(\mathbf{z}) < k\} = \{z : -3z_1 + 7z_2 < -1\}$$

Utilizando las funciones discriminantes lineales del apartado (a)

$$L_2(z) = L_2(z) \longrightarrow \quad -z_1 + 3z_2 - 1 = 2z_1 + 4z_2 - 2$$
$$-3z_1 + 7z_1 = -1$$

d) Dibujar las regiones de clasificación para cada grupo.

$$L_1(z) = L_2(z) \longleftrightarrow z_2 = \frac{1}{3}z_1 + \frac{1}{3}$$
$$L_1(z) = L_3(z) \longleftrightarrow z_2 = \frac{1}{2}z_1 - \frac{1}{2}$$
$$L_2(z) = L_3(z) \longleftrightarrow z_2 = \frac{3}{7}z_1 - \frac{1}{7}$$



3) Dadas tres poblaciones normales con matriz de covarianzas común

$$V = \begin{pmatrix} 6 & 2 \\ 2 & 1 \end{pmatrix}$$

y medias $(0, 1)$, $(1, 0)$ y $(1, 1)$, respectivamente, obtener las funciones discriminantes y el criterio de clasificación.

$$\mu^{(1)} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$\mu^{(2)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$\mu^{(3)} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$V = \begin{pmatrix} 6 & 2 \\ 2 & 1 \end{pmatrix} \longrightarrow V^{-1} = \begin{pmatrix} \frac{1}{2} & -1 \\ -1 & 3 \end{pmatrix}$$

$$L_i(\mathbf{z}) = \left(\mu^{(i)}\right)' V^{-1}\mathbf{z} - \frac{1}{2}\left(\mu^{(i)}\right)' V^{-1}\mu^{(i)}$$

$$L_1(\mathbf{z}) = \begin{pmatrix} 0 & 1 \end{pmatrix}\begin{pmatrix} \frac{1}{2} & -1 \\ -1 & 3 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} - \frac{1}{2}\begin{pmatrix} 0 & 1 \end{pmatrix}\begin{pmatrix} \frac{1}{2} & -1 \\ -1 & 3 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 & 3 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} - \frac{1}{2}\cdot 3 = -x + 3y - \frac{3}{2}$$

$$L_2(\mathbf{z}) = \begin{pmatrix} 1 & 0 \end{pmatrix}\begin{pmatrix} \frac{1}{2} & -1 \\ -1 & 3 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} - \frac{1}{2}\begin{pmatrix} 1 & 0 \end{pmatrix}\begin{pmatrix} \frac{1}{2} & -1 \\ -1 & 3 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & -1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} - \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{2}x - y - \frac{1}{4}$$

$$L_3(\mathbf{z}) = \begin{pmatrix} 1 & 1 \end{pmatrix}\begin{pmatrix} \frac{1}{2} & -1 \\ -1 & 3 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} - \frac{1}{2}\begin{pmatrix} 1 & 1 \end{pmatrix}\begin{pmatrix} \frac{1}{2} & -1 \\ -1 & 3 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -\frac{1}{2} & 2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} - \frac{1}{2}\cdot\frac{3}{2} = -\frac{1}{2}x + 2y - \frac{3}{4}$$

Criterio: z se clasificará en la población donde $L_i(\mathbf{z})$ sea máxima.

4) Dados dos vectores normales bidimensionales con medias $(0,0)$ y $(3,0)$ y matrices de covarianzas

$$V_1 = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \text{ y } V_2 = \begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix},$$

se pide:

a) Calcular las funciones discriminantes cuadráticas

$$V_1^{-1} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}^{-1} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

$$V_2^{-1} = \begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}^{-1} = \begin{pmatrix} 5 & -2 \\ -2 & 1 \end{pmatrix}$$

$$\mu^{(1)} = (0,0)'$$

$$\mu^{(2)} = (3,0)'$$

$$Q_i(\mathbf{z}) = (\mathbf{z} - \mu^{(i)})' V_i^{-1} (\mathbf{z} - \mu^{(i)}) + \log|V_i|$$

$$Q_i^*(\mathbf{z}) = (\mathbf{z} - \mu^{(i)}) V_i^{-1} (\mathbf{z} - \mu^{(i)})$$

Como $|V_i| = 1$, $i = 1, 2$ las dos funciones coinciden.

$$Q_1(\mathbf{z}) = \begin{pmatrix} x-0 & y-0 \end{pmatrix}\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} x-0 \\ y-0 \end{pmatrix} + \cancelto{0}{\log|V_1|} = \begin{pmatrix} 2x+y & x+y \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = 2x^2 + 2xy + y^2$$

$$Q_2(\mathbf{z}) = \begin{pmatrix} x-3 & y-0 \end{pmatrix}\begin{pmatrix} 5 & -2 \\ -2 & 1 \end{pmatrix}\begin{pmatrix} x-3 \\ y-0 \end{pmatrix} + \cancelto{0}{\log|V_2|} = \begin{pmatrix} 5x-2y-15 & -2x+y+6 \end{pmatrix}\begin{pmatrix} x-3 \\ y \end{pmatrix}$$

$$= (5x-2y-15)(x-3) + 2xy + y^2 + 6y = 5x^2 - 15x - 2xy + 6y - 15x + 45 - 2xy + y^2 + 6y$$

$$= 5x^2 + y^2 - 30x + 12y - 4xy + 45$$

Criterio de clasificación: $\mathbf{z}$ se clasifica en la población en la que $Q_i(\mathbf{z})$ mínimo

b) Clasificar a $\mathbf{z} = (1,-4)'$ usando dichas funciones.

$\mathbf{z} = (1,-4)'$ $\quad Q_1(\mathbf{z}) = 2\cdot 1^2 + 2\cdot 1\cdot(-4) + (-4)^2 = 10$

$$Q_2(\mathbf{z}) = 5\cdot 1^2 + (-4)^2 - 30\cdot 1 + 12\cdot(-4) - 4\cdot 1\cdot(-4) + 45 = 4 \quad (*)$$

$\mathbf{z}$ se clasifica en la población (2)

c) Representar gráficamente las regiones de clasificación.

Se clasifica en (1) $\longleftrightarrow Q_1(\mathbf{z}) < Q_2(\mathbf{z})$

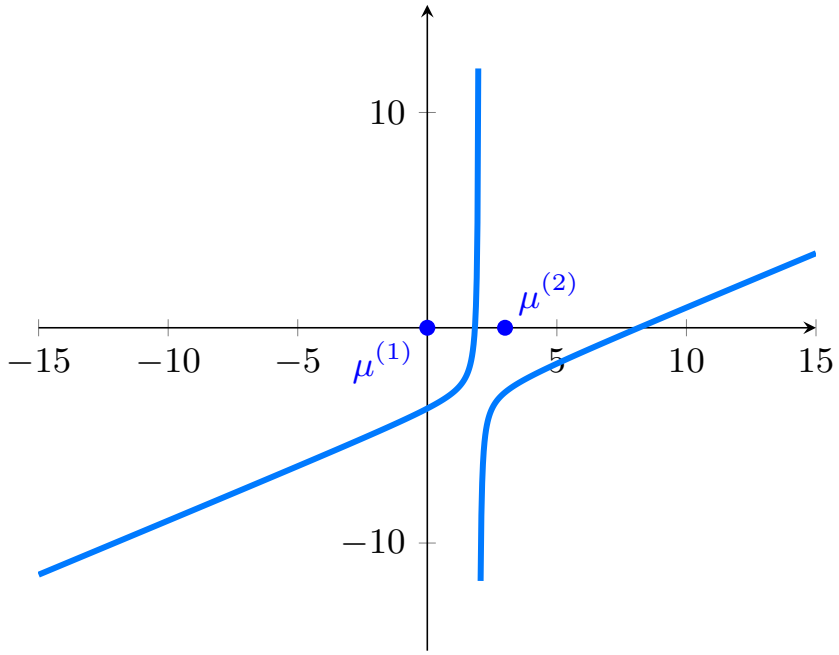
$2x^2 + y^2 + 2xy < 5x^2 + y^2 - 30x + 12y - 4xy + 45 \longleftrightarrow y^2 + 2xy - y^2 - 12y + 4xy < 5x^2 - 30x + 45 - 2x^2$

$\longleftrightarrow 6xy - 12y < 3x^2 - 30x + 45 \xleftrightarrow{\frac{1}{3}} 2xy - 6y < x^2 - 10x + 15 \longleftrightarrow y(-4+2x) < x^2 - 10x + 45$

1) Si $x = 2$, como $0 > -1$, z se clasifica en (2).

2) Si $x > 2$, se clasifica en (1) $\longleftrightarrow y < \dfrac{x^2 - 10x + 15}{-4 + 2x}$.

3) Si $x < 2$, se clasifica en (1) $\longleftrightarrow y > \dfrac{x^2 - 10x + 15}{-4 + 2x}$..



5) Dadas dos poblaciones normales bidimensionales con medias $\mu^{(1)} = (1,0)'$ y $\mu^{(2)} = (0,0)'$ y matrices de covarianzas

$$V_1 = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \text{ y } V_2 = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix},$$

se pide:

a) Obtener las funciones discriminantes cuadráticas y clasificar a $\mathbf{z} = (1,1)'$.

$$V_1 = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \longrightarrow V_1^{-1} = 2 \cdot \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$$

$$V_2 = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \longrightarrow V_2^{-1} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$$

$$\mu_1 = \begin{pmatrix} 1 & 0 \end{pmatrix}' \qquad \mu_2 = \begin{pmatrix} 0 & 0 \end{pmatrix}'$$

$$Q_i(\mathbf{z}) = (\mathbf{z} - \mu^{(i)})' V_i^{-1} (\mathbf{z} - \mu^{(i)}) + \log|V_i|$$

$$Q_1(\mathbf{z}) = \begin{pmatrix} x-1 & y-0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} x-1 \\ y-0 \end{pmatrix} + \log|0.5| = \begin{pmatrix} x-1 & 2y \end{pmatrix} \begin{pmatrix} x-1 \\ y \end{pmatrix} + \log|0.5|$$

$$= (x-1)^2 + 2y^2 + \log|0.5| = x^2 - 2x + 1 + 2y^2 + \log|0.5|$$

$$Q_2(\mathbf{z}) = \begin{pmatrix} x-0 & y-0 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} x-0 \\ y-0 \end{pmatrix} + \cancel{\log(1)}^{\,0} = \begin{pmatrix} x-y & -x+2y \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = x^2 - xy - xy + 2y^2 = x^2 + 2y^2 - 2xy$$

Clasificamos al individuo en $\mathbf{z} = (1,1)'$

$$\begin{cases} Q_1(1,1) = 1.31 \\ Q_2(1,1) = 1^2 + 2 \cdot 1^2 - 2 \cdot 1 \cdot 1 = \boxed{1} \end{cases}$$

Clasificamos al individuo en la población 2.

b) Clasificar a $\mathbf{z}$ usando el criterio de mínima distancia de Mahalanobis y representar las regiones de clasificación con este criterio para cada grupo.

$$Q_i^*(\mathbf{z}) = d_M^2(\mathbf{z}, \mu^{(i)}) = (\mathbf{z} - \mu^{(i)})' V_i^{-1} (\mathbf{z} - \mu^{(i)})$$

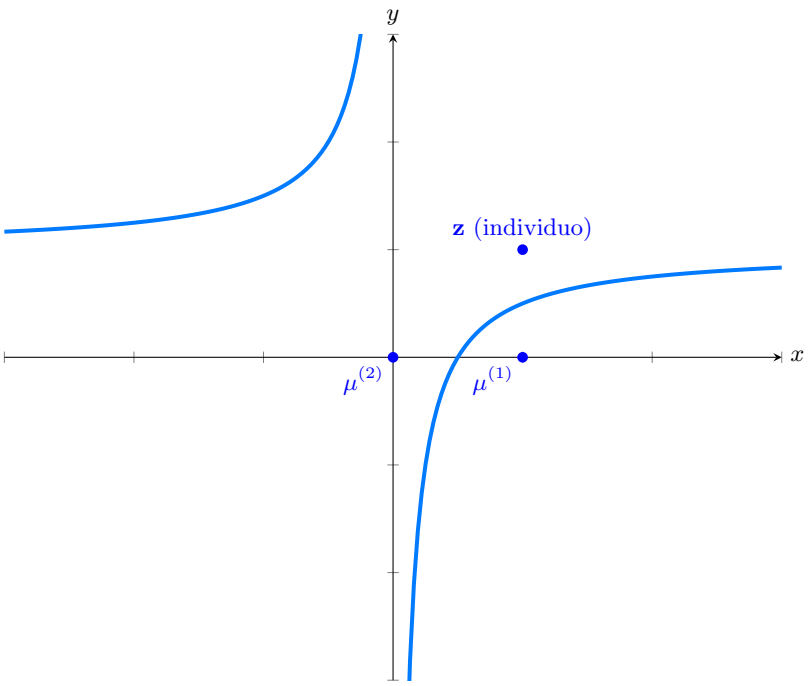
$$Q_1^*(\mathbf{z}) = x^2 + 2y^2 - 2x + 1$$

$$Q_2^*(\mathbf{z}) = x^2 + 2y^2 - 2xy$$

z se clasificará en la población 1 $\longleftrightarrow Q_1^*(\mathbf{z}) < Q_2^*(\mathbf{z}) \longleftrightarrow \cancel{x^2} + \cancel{2y^2} - 2x + 1 < \cancel{x^2} + \cancel{2y^2} - 2xy \xleftrightarrow{\text{Despejamos } y} 2xy < 2x \cdot 1$

Tenemos tres posibilidades:

- Si $x = 0:$ **z** se clasificará en la población 1 $\longleftrightarrow 0 < -1$ No se puede dar, por tanto **z** se clasificará en la población 2
- Si $x > 0:$ **z** se clasificará en la población 1 $\longleftrightarrow \boxed{y < \dfrac{2x-1}{2x}}$
- Si $x < 0:$ **z** se clasificará en la población 1 $\longleftrightarrow \boxed{y > \dfrac{2x-1}{x}}$



Si $x > 0 \qquad y < \dfrac{2x-1}{2x} \longrightarrow \dfrac{2x}{2x} - \dfrac{1}{2x} = 1 - \dfrac{1}{2x}$

Como $x > 0$, si x converge a $0 \longrightarrow 1 - \infty = \boxed{-\infty}$

Si $x < 0 \qquad y > \dfrac{2x-1}{2x} \longrightarrow \dfrac{2x}{2x} - \dfrac{1}{2x} = 1 - \dfrac{1}{2x}$

Como $x > 0$, si x converge a $0 \longrightarrow 1 + \infty = \boxed{+\infty}$

- Interpretación de la gráfica

1) Coordenada x donde habrá una recta vertical donde se produce la convergencia de las exponenciales.
2) Determina en que lado del gráfico se dibuja la exponencial $\begin{cases} x > \longrightarrow \text{ derecha} \\ x < \longrightarrow \text{ izquierda} \end{cases}$
3) Marcará donde converge la exponencial en la recta vertical establecida $\begin{cases} y > \longrightarrow \text{ plano positivo} \\ y < \longrightarrow \text{ plano negativo} \end{cases}$

6) Obtener un criterio de clasificación para dos poblaciones exponenciales unidimensionales con medias distintas usando máxima verosimilitud. Clasificar a $z = 1.5$ entre dos poblaciones exponenciales con medias 2 y 1. (Indicación: La función de densidad de la distribución exponencial es $f(x) = \dfrac{e^{-\frac{x}{\mu}}}{\mu}$ para $x \geq 0$).

$$f(x) = \begin{cases} \dfrac{1}{\mu} e^{-\frac{x}{\mu}} & \text{si } x > 0 \\ 0 & \text{en otro caso} \end{cases}$$

Población (1): media= $E[X^{(1)}] = \mu_1$; f_1 función de densidad

Población (2): media= $E[X^{(2)}] = \mu_2$; f_2 función de densidad

Criterio de máxima verosimilitud:

Se clasificará **z** en (1)$\longleftrightarrow f_1(\mathbf{z}) > f_2(\mathbf{z}) \longleftrightarrow \dfrac{1}{\mu_1} e^{-\frac{\mathbf{z}}{\mu_1}} > \dfrac{1}{\mu_2} e^{-\frac{\mathbf{z}}{\mu_2}} \longleftrightarrow e^{-\frac{\mathbf{z}}{\mu_1}} e^{\frac{\mathbf{z}}{\mu_2}} > \dfrac{\mu_1}{\mu_2} \longleftrightarrow e^{\mathbf{z}\left(\frac{1}{\mu_2} - \frac{1}{\mu_1}\right)} > \dfrac{\mu_1}{\mu_2} \longleftrightarrow \mathbf{z}\left(\dfrac{1}{\mu_2} - \dfrac{1}{\mu_1}\right) > \ln\left(\dfrac{\mu_1}{\mu_2}\right)$

Si $\mu_1 > \mu_2$, se clasifica en (1) $\longleftrightarrow \mathbf{z} > \dfrac{1}{\left(\frac{1}{\mu_2} - \frac{1}{\mu_1}\right)} \ln\left(\dfrac{\mu_1}{\mu_2}\right)$

Si $\mu_1 < \mu_2$, se clasifica en (1) $\longleftrightarrow \mathbf{z} < \dfrac{1}{\left(\frac{1}{\mu_2} - \frac{1}{\mu_1}\right)} \ln\left(\dfrac{\mu_1}{\mu_2}\right)$

$\mu_1 = 2,\ \mu_2 = 1$ $\quad \begin{cases} f_1(1.5) = 0.2361 & (*) \text{ Se clasificará en la población (1)} \\ f_2(1.5) = 0.2231 \end{cases}$

$\mathbf{z} = 1.5$

Ejercicio de clase (22/04/2024): Sea (X, Y) un vector aleatorio con función de densidad

$$f(x,y) = \begin{cases} \dfrac{2x+2y}{5} & 0 < x+y < 2, 0 < y < 1 \\ 0 & \text{en otro caso} \end{cases}$$

a) Calcular la recta de regresión de X sobre Y y obtener una medida de la bondad del ajuste realizado.

Recta de regresión de X sobre Y:

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

$E[X] = \iint x f(x,y)\,\mathrm{d}x\,\mathrm{d}y$

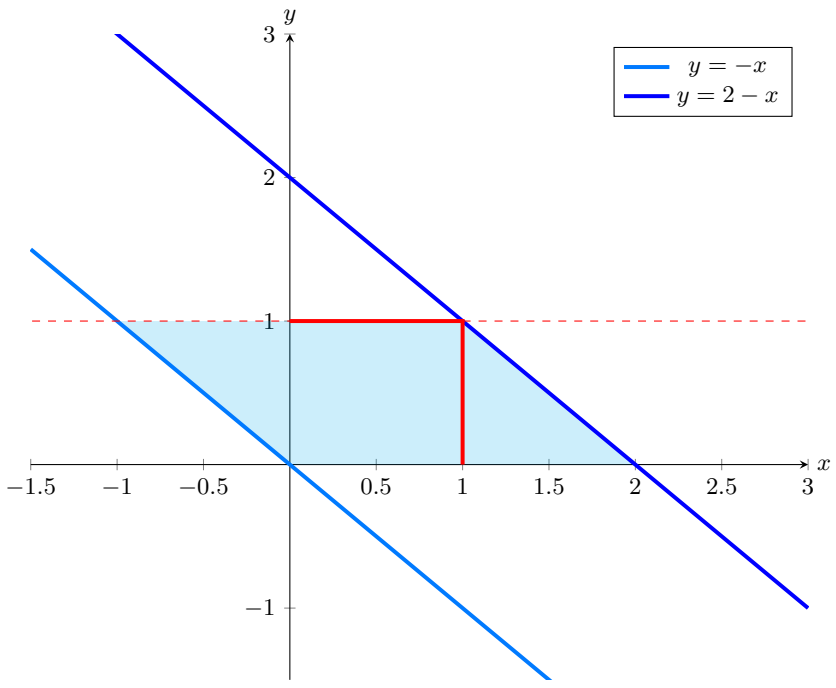
$E[Y] = \iint y f(x,y)\,\mathrm{d}x\,\mathrm{d}y$

$\mathrm{Var}[Y] = E[Y^2] - (E[Y])^2$

$E[Y^2] = \iint y^2 f(x,y)\,\mathrm{d}x\,\mathrm{d}y$

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

$E[X \cdot Y] = \iint x \cdot y \cdot f(x,y)\,\mathrm{d}x\,\mathrm{d}y$



b) Calcular la curva de regresión de X sobre Y.

c) Predecir el valor de X cuando Y toma el valor $\frac{1}{2}$ mediante la curva y la recta, y dar un valor del error de predicción en cada caso.

Tema 6: Análisis Cluster

6.1) Introducción

6.1.1) Objetivo

Objetivo: cómo agrupar observaciones estableciendo grupos (o clusters) con las más similares.

Aprendizaje supervisado: en la muestra se indica a qué grupo pertenece cada observación.

- Regresión Logística.
- Análisis Discriminante.

Aprendizaje no supervisado (o automático): en este caso no disponemos de una muestra inicial donde se indique a qué grupo pertenece cada observación. De hecho, en algunas ocasiones podemos decidir cuántos grupos queremos establecer.

- Análisis Cluster

6.1.2) Contexto

Dispondremos de una muestra (o población) de n individuos (objetos) en los que hemos medido k variables numéricas $(X_1, \ldots, X_k)$.

Sin embargo, en este caso, no dispondremos de una variable Y que nos diga a qué grupo (población) pertence cada observación.

Incluso, en algunos casos, no sabremos ni siquiera el número de grupos.

De hecho, lo que haremos será determinar los valores de Y que nos asigne los grupos que minimicen una función costo adecuada.

Para ello tendremos que utilizar una función distancia que nos mida cómo de similares son dos observaciones (individuos).

La elección de esta distancia es muy importante y la solución final dependerá de la distancia elegida.

6.2) Distancias entre individuos

6.2.1) La distancia Euclídea

La distancia más popular es la distancia Euclídea, definida como

$$d_E(\mathbf{x}, \mathbf{c}) = \sqrt{(\mathbf{x}-\mathbf{c})'(\mathbf{x}-\mathbf{c})} = \sqrt{\sum_{j=1}^{k}(x_j - c_j)^2}$$

para todo $\mathbf{x}, \mathbf{c} \in \mathbb{R}^k$ (vectores columna).

En nuestro contexto, habitualmente $\mathbf{x} = (x_1, \ldots, x_k)'$ representará un individuo y $\mathbf{c} = (c_1, \ldots, c_k)'$ el centroide de un grupo.

En R se puede computar como

```
dE <- function(x, y) sqrt(sum((x - y)*(x - y)))
```

Por ejemplo, para $x = (0,0)'$ e $y = (1,1)'$

```
x <- c(0,0)
y <- c(1,1)
dE(x,y)
```

```
## [1] 1.414214
```

6.2.2) La distancia de Mahalanobis

Otra opción es la distancia de Mahalanobis que usa le métrica de los datos, definida como

$$d_M(\mathbf{x},\mathbf{c}) = \sqrt{(\mathbf{x}-\mathbf{c})' V^{-1} (\mathbf{x}-\mathbf{c})},$$

donde $V = \text{Cov}(X_1, \ldots, X_m)$.

El principal problema es que si hay grupos, esta matriz puede ser distinta en cada grupo.

Incluso, aunque supongamos que todos los grupos tienen la misma matriz de covarianzas, estos tendrán medias distintas y, como desconocemos los grupos, no podemos estimar estimar V (como hacíamos en el Ánalisis Discriminante).

Una solución es suponer inicialmente que todos los individuos están en un mismo grupo (población) y calcular (estimar) la media y la covarianzas en ella.

En R se puede calcular la función `mahalanobis(x, y, V)`, que proporciona el cuadro de esta distancia, para

$$V = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix},$$

$x = (0,0)'$ e $y = (1,1)'$.

```
V <- matrix(c(1, 1/2,
              1/2, 2), nrow = 2, ncol = 2, byrow = TRUE)
x <- c(0,0)
y <- c(1,1)
mahalanobis(x, y, V)
```

```
## [1] 1.333333
```

O bien, como

```
dM <- function(x, y, V) sqrt(sum(t(x - y ) %*% solve(V) %*% (x-y)))
dM(x, y, V)
```

```
## [1] 1.154701
```

```
## Si hacemos el cuadrado
dM(x, y, V)^2
```

```
## [1] 1.333333
```

Obviamente, si $V = I$ (matriz identidad), se obtiene la distancia Euclídea que, por lo tanto, representará a variables aleatorias independientes con varianza uno.

En otros casos, la distancia de Mahalanobis tendrá en cuenta las varianzas de las variables y sus covarianzas (correlaciones o dependencia).

Las circunferencias (elipsoides) obtenidas con $d_V(\mathbf{x},\mu,V) = \text{cte.}$ coincidirán con los conjuntos de nivel de la distribución normal multivariante $\mathcal{N}_k(\mu, V)$ cuya función de densidad es

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

- De esta forma, bajo este modelo y si conocemos V, el individuo con medidas $\mathbf{x}$ se asignará al grupo en donde sea más verosímil, es decir, donde$f_i(x)$ sea máxima, siendo f_i, la densidad $\mathcal{N}_k(\mu_i, V)$ (tal y como hacíamos en Análisis Discriminante).

Ahora el problema es que no sabemos cómo estimar μ_i y V y tampoco sabemos si hay una matriz de covarianzas V común.

6.2.3) Otras distancias interesantes

La distancia absoluta (Manhattan, de ciudad o geométrica del taxista)

$$d_A(\mathbf{x},\mathbf{c}) = \sum_{j=1}^{k} |x_j - c_j|$$

(que usa las cuadrículas como caminos).

La distancia L_s

$$d_s(\mathbf{x},\mathbf{c}) = \left(\sum_{j=1}^{k} (x_j - c_j)^s\right)^{\frac{1}{s}}$$

para $s > 0$.

La distancia de Pearson

$$d_P(\mathbf{x},\mathbf{c}) = \sqrt{\sum_{j=1}^{k} \left(\frac{x_j - c_j}{\sigma_j}\right)^2}$$

donde σ_i es la desviación típica de X_i, $i = 1, \ldots, k$.

- Este último caso es equivalente a estandarizar los datos usando $Z_i = \frac{X_i}{\sigma_i}$ o $Z_i = \frac{X_i - \mu_i}{\sigma_i}$ lo que nos asegura que las variables tendrán magnitudes similares aunque se usen unidades diferentes en ellas (esto no ocurre en la distancia Euclídea).
- El principal problema es que desconocemos σ_i y μ_i que tendrán que ser estimados usando todos los datos (sin grupos).
- Obviamente, es equivalente a usar la distancia Euclídea con los datos estandárizados.
- La distancia no dependerá de las unidades usadas en cada variable (es invariante por cambio de escala).

6.3) Distancia de individuos a grupos y distancias entre grupos

6.3.1) Distancias de individuos a grupos

Además de definir las distancias entre individuos, también tendremos que definir distancias de individuos a grupos o distancias entre grupos, lo que nos llevará a definir diversas funciones coste que determinarán diferentes soluciones finales.

- Estas vendrán determinadas por el problema que queremos resolver.

Por ejemplo, si queremos calcular la distancia de un individuo $\mathbf{x}$ a un grupo $\{\mathbf{z}_i : i \in G\}$ formado por $m = |G|$ individuos podemos definir las distancias siguientes:

$$d_1(\mathbf{x}, G) \coloneqq d(\mathbf{x}, \mathbf{C}), \quad \mathbf{C} = \frac{1}{|G|}\sum_{i\in G} \mathbf{z}_i$$

$$d_2(\mathbf{x}, G) \coloneqq \min_{i\in G} d(\mathbf{x}, \mathbf{z}_i),$$

$$d_3(\mathbf{x}, G) \coloneqq \max_{i\in G} d(\mathbf{x}, \mathbf{z}_i),$$

$$d_4(\mathbf{x}, G) \coloneqq \sum_{i\in G} d(\mathbf{x}, \mathbf{z}_i),$$

$$d_5(\mathbf{x}, G) \coloneqq \sum_{i\in G} d^2(\mathbf{x}, \mathbf{z}_i),$$

donde d es una distancia entre individuos.

Otra opción interesante es calcular (o estimar) una función de densidad para los individuos de un mismo grupo y calcular las distancias como

$$d(\mathbf{x}, G_j) = 1 - \frac{f_j(\mathbf{x})}{f_1(\mathbf{x}) + \cdots + f_m(\mathbf{x})}.$$

Análogamente, para las distancias entre grupos se pueden usar:

$$D_1(G-1,G_2)=d(C_1,C_2),\quad C_j=\frac{1}{|G_j|}\sum_{i\in G_j}\mathbf{z}_i,\quad j=1,2$$

$$D_2(G_1,G_2)=\min_{i\in G_1,j\in G_2} d(\mathbf{z}_i,\mathbf{z}_j),$$

$$D_3(G_1,G_2)=\max_{i\in G_1,j\in G_2} d(\mathbf{z}_i,\mathbf{z}_j),$$

$$D_4(G_1,G_2)=\frac{1}{|G_1|\cdot|G_2|}\sum_{i\in G_1,j\in G_2} d(\mathbf{z}_i,\mathbf{z}_j),$$

$$D_5(G_1,G_2)=\frac{1}{|G_1|\cdot|G_2|}\sum_{i\in G_1,j\in G_2} d^2(\mathbf{z}_i,\mathbf{z}_j).$$

En d_1 o en D_1 podemos utilizar otros centroides C_1 y C_2 distintos de la medida de cada grupo.

Estas distancias entre grupos nos permitirán representar sus distancias y, posteriormente establecer a partir de qué nivel uniremos los grupos formando los gráficos denominados dendogramas.

Finalmente debemos definir una función costo que trataremos de minimizar para obtener la solución óptima de ese problema.

6.3.2) Función costo

Supongamos que asignamos los n individuos a un grupo mediante una variable Y que nos indicará con $y_i=j$ que el individuo i se asigna al grupo j.

Podemos definir la función costo

$$J(y)=\sum_j\sum_{i:y_i=j} d(\mathbf{x}_i,G_j),$$

donde $\sum_{j:y_i=j} 1=1$ para todo i (cada elemento se asigna a un único grupo).

También se pueden usar distancias al cuadrado

$$J^*(y)=\sum_j\sum_{i:y_i=j} d^2(\mathbf{x}_i,G_j).$$

En estos métodos, tenemos que fijar un número máximo de grupos ya que si no, la solución óptima será tener n grupos (uno para cada elemento).

Otra opción podría ser maximizar la sima total de las distancias entre grupos para la clasificación y:

$$D(y)=\sum_{i<j} D(G_i,G_j).$$

Todas estas opciones nos llevarán a problemas diferentes que tendrán que resolverse (cuando sea posible) usando sus técnicas específicas (la mayoría de Investigación Operativa).

6.4) Métodos cluster

6.4.1) Clasificación

Estos métodos se pueden dividir en dos grandes grupos:

- Los métodos jerárquicos: Parten de la idea de juntar las unidades (individuos o grupos) más similares (cercanas).
- Los métodos no jerárquicos: Establecen un determinado número de grupos y se irá asignado cada individuo al grupo más cercano.

Solamente veremos un método de cada tipo.

6.5) Método no jerárquico de las K-medias

El método de las K-medias (K-means) es sin duda el método no jerárquico más popular.

Habitualmente usa la distancia Euclídea con los datos sin estandarizar (cuando tienen escalas similares) o estandarizados (distancia de Pearson, cuando tienen escalas diferentes) pero se puede aplicar a otras distancias.

En este caso tenemos que fijar un número de grupos predeterminado K con $1 < K \leq \frac{n}{2}$.

Posteriormente podremos aumentar o disminuir K según la solución obtenida.

- K es el número de grupos.
- k es el número de variables.
- n es el número de observaciones.
- Estos números pueden ser diferentes.

6.5.1) Algoritmo del método de las K-medias

- Paso 0: Determinar K centroides $C_1^0, \ldots, C_K^0 \in \mathbb{R}^k$ al azar.
- Paso 1: En la iteración m, formar el grupo G_j^m con las observaciones que están más cercanas al centroide C_j^{m-1} para $j = 1, \ldots, K$.
- Paso 2: Calcular el centroide C_j^m del grupo G_k^m definido como el punto que minimiza

$$\sum_{i \in G_j^m} d(\mathbf{x}_i, C_j^m),$$

  o considerando las distancias al cuadrado

$$\sum_{i \in G_j^m} d^2(\mathbf{x}_i, C_j^m),$$

  para $j = 1, \ldots, K$.
- Paso 3: Repetir pasos 1 y 2 hasta que no se produzcan cambios en los grupos del paso o hasta que se haya iterado un número determinado de veces.

Si usamos la distancia Euclídea y el error cuadrático, los centroides del paso 2 serán las medias aritméticas de los datos de cada grupo ya que si queremos minimizar

$$\min_P \sum_{j \in G} d^2(\mathbf{O}_j, P)$$

para un grupo G, tenemos que

$$\begin{aligned}
\sum_{j \in G} d^2(\mathbf{O}_j, P) &= \sum_{j \in G} (\mathbf{O}_j - P)'(\mathbf{O}_j - P) \\
&= \sum_{j \in G} (\mathbf{O}_j - \overline{\mathbf{O}}_G + \overline{\mathbf{O}}_G - P)'(\mathbf{O}_j - \overline{\mathbf{O}}_G + \overline{\mathbf{O}}_G - P) \\
&= |G|(\overline{\mathbf{O}}_G - P)'(\overline{\mathbf{O}}_G - P) + \sum_{j \in G} (\mathbf{O}_j - \overline{\mathbf{O}}_G)'(\mathbf{O}_j - \overline{\mathbf{O}}_G) + 2 \sum_{j \in G} (\overline{\mathbf{O}}_G - P)'(\mathbf{O}_j - \overline{\mathbf{O}}_G) \\
\sum_{j \in G} d^2(\mathbf{O}_j, P) &= |G|(\overline{\mathbf{O}}_G - P)'(\overline{\mathbf{O}}_G - P) + \sum_{j \in G} (\mathbf{O}_j - \overline{\mathbf{O}}_G)'(\mathbf{O}_j - \overline{\mathbf{O}}_G) + 2(\overline{\mathbf{O}}_G - P)' \sum_{j \in G} (\mathbf{O}_j - \overline{\mathbf{O}}_G) \\
&= |G|(\overline{\mathbf{O}}_G - P)'(\overline{\mathbf{O}}_G - P) + \sum_{j \in G} (\mathbf{O}_j - \overline{\mathbf{O}}_G)'(\mathbf{O}_j - \overline{\mathbf{O}}_G),
\end{aligned}$$

donde $\overline{\mathbf{O}}_G = \frac{1}{|G|} \sum_{j \in G} \mathbf{O}_j$ es la media del grupo G.

En la expresión final, el segundo sumando es constante (no depende de P) y el mínimo del primer sumando sumando se alcanza con $P = \overline{\mathbf{O}}_G$ (ya que es la distancia entre esos dos puntos).

De esta forma el paso 2 es inmediato y en el paso 1 simplemente calculamos las distancias a estos nuevos K centroides (medias) asignando cada individuo al grupo del centroide más cercano (distancia d_1).

El nombre `k-means` proviene de esta propiedad.

Además, si usamos como función de coste las dadas previamente y hay cambios en los grupos, esta función es estrictamente decrecientes en el paso 1 ya que hay al menos un `objeto` cuya distancia al grupo (centroide) ha disminuido.

Al recalcular los centroides en el paso 2 las sumas de estas distancias en los grupos disminuirán aún más o se quedarán iguales a las del paso 1.

Como las opciones del paso 1 son finitas, este algoritmo conducirá hasta una solución óptima local en un número finito de pasos, que puede ser muy grande.

- Para evitar este problema podemos aplicar el algoritmo varias veces con centroides iniciales diferentes y comparar las soluciones óptimas finales de cada algoritmo.

Veremos que con unos pocos pasos podemos obtener soluciones muy buenas.

6.5.2) Un ejemplo sencillo

Mostramos con un ejemplo sencillo el funcionamiento del algoritmo.

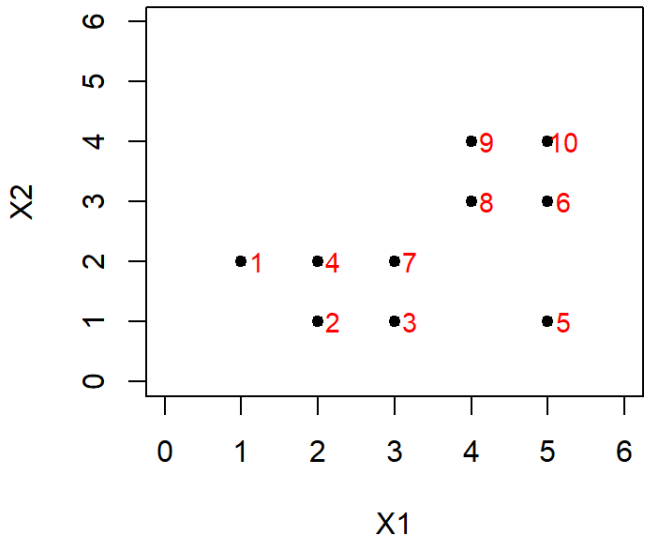
Para ello usaremos los datos analizados previamente con regresión logística pero ahora supondremos que no conocemos los grupos de esa muestra.

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

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

Individuos sin agrupamiento inicial.

```
X1 <- c(1, 2, 3, 2, 5, 5, 3, 4, 4, 5)
X2 <- c(2, 1, 1, 2, 1, 3, 2, 3, 4, 4)
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, 1:length(X1), cex = 0.9, col = "red")
```



Observamos que tienen unidades similares (por lo que podremos usar la distancia Euclídea) y que parecen formar dos grupos diferentes.

El objetivo es determinar la variable Y que nos asigne cada individuo a un grupo.

En tabla anterior, hemos dejado en blanco la columna de la variable Y para señalar que en este caso no tenemos una muestra de entrenamiento y, por lo tanto, no podremos saber cuál es la solución óptima (que mejor clasifique a los individuos).

- Análisis no supervisado (o automático).

Para aplicar el algoritmo con $K = 2$ medias (grupos) elegimos dos centroides al azar (dentro de la zona donde están los individuos).

Para ello usamos la instrucción `runif(2,0,6)` (fijando previamente la semilla con `set.seed`).

Primer centroide en el paso 0:

```
set.seed(123124)
C1_0 = runif(2, 0, 6)
C1_0
```

```
## [1] 3.101323 1.401716
```

Segundo centroide en el paso 0:

```
set.seed(123121)
C2_0 = runif(2, 0, 6)
C2_0
```

```
## [1] 2.2950230 0.3726236
```

Otra opción sería usar dos de esos puntos al azar.

Con los centroides obtenidos, $C_1^0 = (3.101323, 1.401716)$ y $C_2^0 = (2.2950230, 0.3726236)$, obtenemos las distancias y agrupaciones siguientes:

```
library("dplyr")
df = data.frame(X1, X2) %>%
  mutate(d_C1 = sqrt((X1 - C1_0[1])^2 + (X2 - C1_0[2])^2),
         d_C2 = sqrt((X1 - C2_0[1])^2 + (X2 - C2_0[2])^2),
```

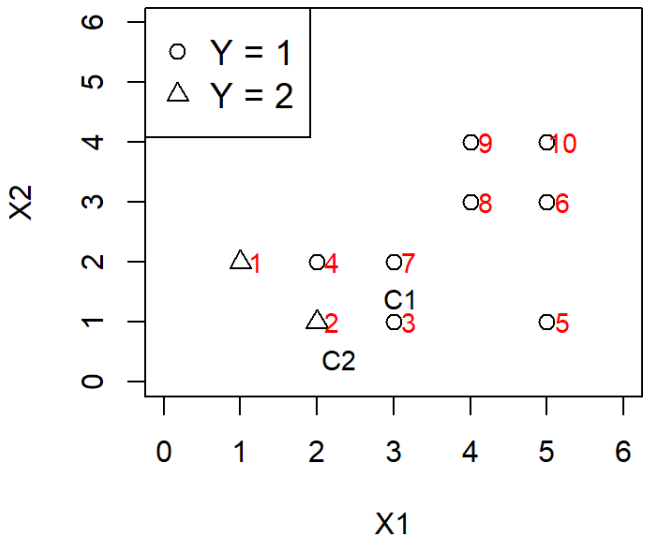
```
        Y = ifelse(d_C1 <  d_C2, 1, 2))
```

Individuo i	X_1	X_2	$d(\mathbf{X}, C_1^0)$	$d(\mathbf{X}, C_2^0)$	Y
1	1	2	2.1848343	2.0797689	2
2	2	1	1.1723001	0.6932819	2
3	3	1	0.4142971	0.9437127	1
4	2	2	1.2533377	1.6539022	1
5	5	1	1.9407090	2.7767790	1
6	5	3	2.4818314	3.7709425	1
7	3	2	0.6068031	1.7735125	1
8	4	3	1.8336119	3.1321005	1
9	4	4	2.7493091	4.0080926	1
10	5	4	3.2180825	4.529044	1

Cada individuo se asigna al grupo más cercano (midiendo su distancia a cada centroide).

Individuos agrupados en el primer paso del algoritmo K-means con los centroides iniciales (negro).

```
Y <- df$Y
plot(X1, X2, xlab = "X1", ylab = "X2", pch = as.integer(Y),
     xlim = c(0,6), ylim = c(0,6), cex = 1.2)
text(X1 + 0.2, X2, 1:length(X1), cex = 0.9, col = "red")
text(C1_0[1], C1_0[2], "C1", cex = 0.9, col = "black")
text(C2_0[1], C2_0[2], "C2", cex = 0.9, col = "black")
legend('topleft', legend = c('Y = 1','Y = 2'),
       pch = 1:2, cex = 1.2)
```



En el siguiente paso, calculamos los nuevos centroides, que son las medias de los individuos de cada grupo.

```
centroides = df %>%
  group_by(Y) %>%
  summarise(C_x1 = mean(X1),
            C_x2 = mean(X2)) %>%
```

```
  ungroup
C1_1 = c(centroides$C_x1[1], centroides$C_x2[1])
C1_1
```

```
## [1] 3.875 2.500
```

```
C2_1=  c(centroides$C_x1[2], centroides$C_x2[2])
C2_1
```

```
## [1] 1.5 1.5
```

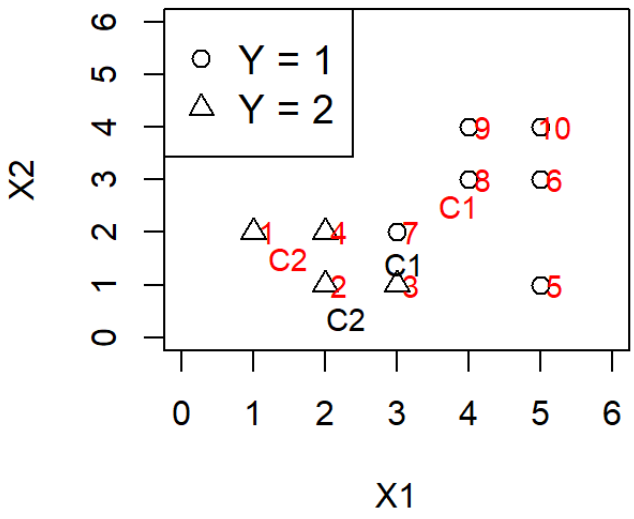
Estos centroides son:

$C_1^1 = (3.875, 2.5)$

$C_2^1 = (1.5, 1.5)$.

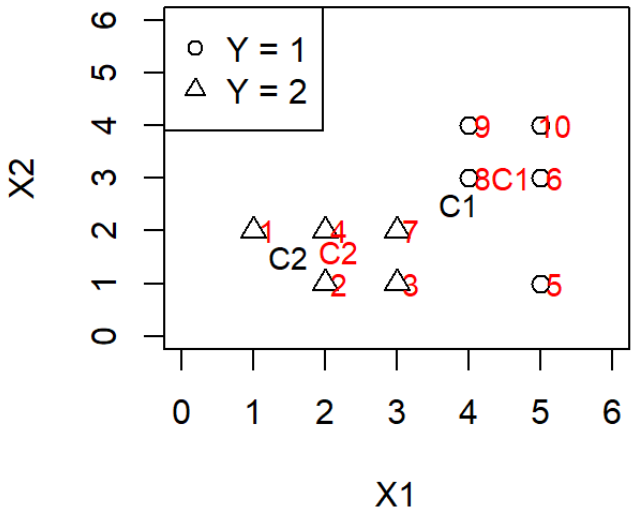
Individuos agrupados en el primer y segundo paso del algoritmo K-means con los centroides iniciales (negro) y los nuevos (rojo).

```
df = df %>%
  mutate(d_C1 = sqrt((X1 - C1_1[1])^2 + (X2 - C1_1[2])^2),
         d_C2 = sqrt((X1 - C2_1[1])^2 + (X2 - C2_1[2])^2),
         Y = ifelse(d_C1 < d_C2,
                    1,
                    2))
Y <- df$Y
plot(X1, X2, xlab = "X1", ylab = "X2", pch = as.integer(Y),
     xlim = c(0,6), ylim = c(0,6), cex = 1.2)
text(X1 + 0.2, X2, 1:length(X1), cex = 0.9, col = "red")
text(C1_0[1], C1_0[2], "C1", cex = 0.9, col = "black")
text(C2_0[1], C2_0[2], "C2", cex = 0.9, col = "black")
text(C1_1[1], C1_1[2], "C1", cex = 0.9, col = "red")
text(C2_1[1], C2_1[2], "C2", cex = 0.9, col = "red")
legend('topleft', legend = c('Y = 1','Y = 2'), pch = 1:2, cex = 1.2)
```



Repetimos los cálculos con los nuevos centroides. El reparto del paso uno conduce a los mismos grupos en la iteración anterior y el algoritmo se detiene, obteniendo los centroides finales: $C_1^2 = (4.6, 3.0)$ y $C_2^2 = (2.2, 1.6)$.

```
df = df %>%
  mutate(d_C1 = sqrt((X1 - C1_1[1])^2 + (X2 - C1_1[1])^2),
         d_C2 = sqrt((X1 - C2_1[1])^2 + (X2 - C2_1[2])^2),
         Y = ifelse(d_C1 < d_C2,
                    1,
                    2))

centroides = df %>%
  group_by(Y) %>%
  summarise(C_x1 = mean(X1),
            C_x2 = mean(X2))
C1_2 = c(centroides$C_x1[1], centroides$C_x2[1])
C2_2 =  c(centroides$C_x1[2], centroides$C_x2[2])
df = df %>%
  mutate(d_C1 = sqrt((X1 - C1_2[1])^2 + (X2 - C1_2[2])^2),
         d_C2 = sqrt((X1 - C2_2[1])^2 + (X2 - C2_2[2])^2),
         Y = ifelse(d_C1 < d_C2,
                    1,
                    2))
Y <- df$Y
plot(X1, X2, xlab = "X1", ylab = "X2", pch = as.integer(Y),
     xlim = c(0,6), ylim = c(0,6), cex = 1.2)
text(X1 + 0.2, X2, 1:length(X1), cex = 0.9, col = "red")
text(C1_1[1], C1_1[2], "C1", cex = 0.9, col = "black")
text(C2_1[1], C2_1[2], "C2", cex = 0.9, col = "black")
text(C1_2[1], C1_2[2], "C1", cex = 0.9, col = "red")
text(C2_2[1], C2_2[2], "C2", cex = 0.9, col = "red")
legend('topleft', legend = c('Y = 1','Y = 2'), pch = 1:2, cex = 1)
```



Cuando los grupos no varían en dos iteraciones seguidas los centroides coinciden y el algoritmo se detiene.

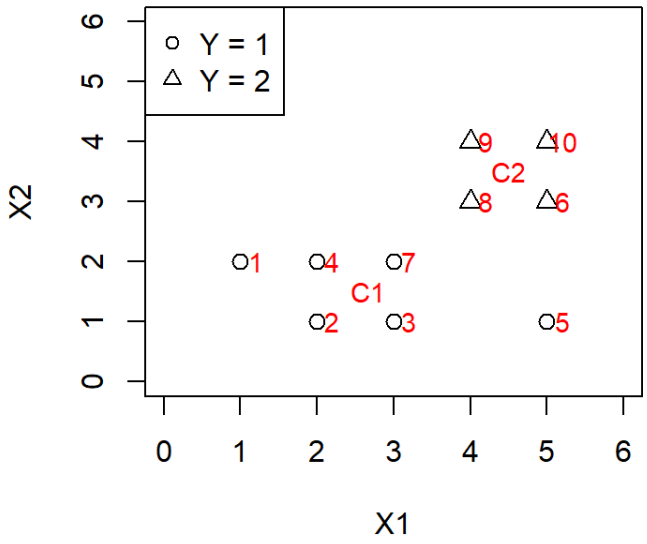
Problema: este algoritmo puede depender de los valores iniciales.

Presentamos la agrupación de los individuos con otros centroides iniciales.

Centroides iniciales: $C_1^0 = (2.482099, 2.270985)$ y $C_2^0 = (3.851084, 5.344700)$.

Centroides finales: $C_1 = (2.6667, 1.5)$ y $C_2 = (4.5, 3.5)$.

```
K = 2
## Paso 0
C = matrix(NA, ncol = 2, nrow = K)

C[1, ] = c(2.482099,2.270985)
C[2, ] = c(3.851084, 5.344700)

clusters = matrix(NA, ncol = 2, nrow = nrow(df))
clusters[, 1] = 1
clusters[, 2] = 2

i = 2
while(sum(clusters[, 1] == clusters[, 2]) != nrow(df)) {
## Paso 1
df = df %>%
  mutate(d_C1 = sqrt((X1 - C[1, 1])^2 + (X2 - C[1, 2])^2),
         d_C2 = sqrt((X1 - C[2, 1])^2 + (X2 - C[2, 2])^2),
         Y = ifelse(d_C1 < d_C2,
                    1,
                    2))
clusters[, 1] = df$Y
## Paso 2
centroides = df %>%
  group_by(Y) %>%
  summarise(C_x1 = mean(X1),
            C_x2 = mean(X2))
C[1, ] = c(centroides$C_x1[1], centroides$C_x2[1])
C[2, ] =  c(centroides$C_x1[2], centroides$C_x2[2])
df = df %>%
  mutate(d_C1 = sqrt((X1 - C[1, 1])^2 + (X2 - C[1, 2])^2),
         d_C2 = sqrt((X1 - C[2, 1])^2 + (X2 - C[2, 2])^2),
         Y = ifelse(d_C1 < d_C2,
                    1,
                    2))
clusters[, 2] = df$Y
i = i + 1
}

Y <- df$Y

plot(X1, X2, xlab = "X1", ylab = "X2", pch = as.integer(Y),
     xlim = c(0,6), ylim = c(0,6), cex = 1.2)
text(X1 + 0.2, X2, 1:length(X1), cex = 0.9, col = "red")
text(C[1, 1], C[1, 2], "C1", cex = 0.9, col = "red")
text(C[2, 1], C[2, 2], "C2", cex = 0.9, col = "red")
legend('topleft', legend = c('Y = 1','Y = 2'), pch = 1:2, cex = 1)
```

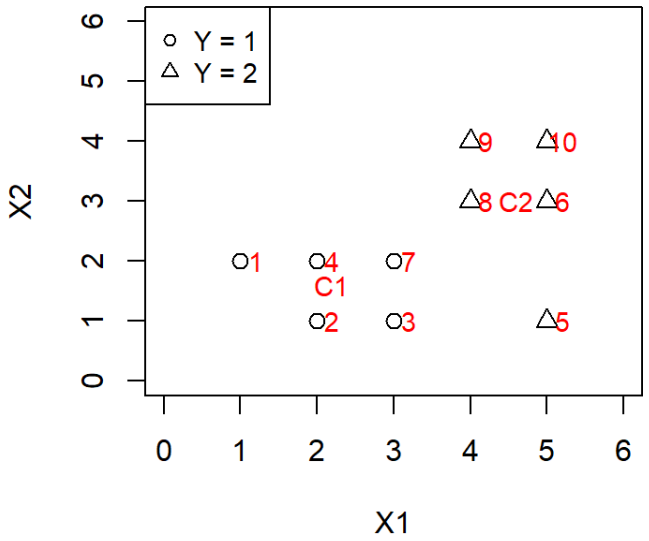


Centroides iniciales: $C_1^0 = (2, 3)$ y $C_2^0 = (5, 3)$.

Centroides finales: $C_1 = (2.2, 1.6)$ y $C_2 = (4.6, 3.0)$

```r
library("dplyr")
K = 2
## Paso 0
C = matrix(NA, ncol = 2, nrow = K)

C[1, ] = c(2, 3)
C[2, ] = c(5, 3)

clusters = matrix(NA, ncol = 2, nrow = nrow(df))
clusters[, 1] = 1
clusters[, 2] = 2

i = 2
while(sum(clusters[, 1] == clusters[, 2]) != nrow(df)) {
## Paso 1
df = df %>%
  mutate(d_C1 = sqrt((X1 - C[1, 1])^2 + (X2 - C[1, 2])^2),
         d_C2 = sqrt((X1 - C[2, 1])^2 + (X2 - C[2, 2])^2),
         Y = ifelse(d_C1 < d_C2,
                    1,
                    2))
clusters[, 1] = df$Y
## Paso 2
centroides = df %>%
  group_by(Y) %>%
  summarise(C_x1 = mean(X1),
            C_x2 = mean(X2))
C[1, ] = c(centroides$C_x1[1], centroides$C_x2[1])
C[2, ] =  c(centroides$C_x1[2], centroides$C_x2[2])
df = df %>%
```

```
  mutate(d_C1 = sqrt((X1 - C[1, 1])^2 + (X2 - C[1, 2])^2),
         d_C2 = sqrt((X1 - C[2, 1])^2 + (X2 - C[2, 2])^2),
         Y = ifelse(d_C1 < d_C2,
                    1,
                    2))
clusters[, 2] = df$Y
i = i + 1
}

Y <- df$Y

plot(X1, X2, xlab = "X1", ylab = "X2", pch = as.integer(Y),
     xlim = c(0,6), ylim = c(0,6), cex = 1.2)
text(X1 + 0.2, X2, 1:length(X1), cex = 0.9, col = "red")
text(C[1, 1], C[1, 2], "C1", cex = 0.9, col = "red")
text(C[2, 1], C[2, 2], "C2", cex = 0.9, col = "red")
legend('topleft', legend = c('Y = 1','Y = 2'), pch = 1:2, cex = 0.9)
```



6.5.3) ¿Cómo comparar distintas soluciones?

Para comparar las soluciones podemos utilizar diversas medidas.

Por ejemplo podíamos usar la función de costo J (con distancias Euclídeas).

```
## función costo J(y)
J <- function(df, C) {
  J= df %>%
    group_by(Y) %>%
    summarise(d = ifelse(Y ==1,
                         sqrt((X1 - C[1, 1])^2 + (X2 - C[1, 2])^2),
                         sqrt((X1 - C[2, 1])^2 + (X2 - C[2, 2])^2))) %>%
    ungroup
  return(sum(J$d))
}
```

Para la solución $C_1 = (2.66667, 1.5)$ y $C_2 = (4.5, 3.5)$ (***sol1***):

```
C = matrix(c(2.666667, 1.5,
             4.5, 3.5), nrow = 2, ncol =2, byrow = TRUE)
df = df %>%
  mutate(d_C1 = sqrt((X1 - C[1, 1])^2 + (X2 - C[1, 2])^2),
         d_C2 = sqrt((X1 - C[2, 1])^2 + (X2 - C[2, 2])^2),
         Y = ifelse(d_C1 < d_C2,
                    1,
                    2))
J(df, C)
```

```
## [1] 3.823299
```

$J(y_{sol1}) = 9.823299$.

Para la solución $C_1 = (2.2, 1.6)$ y $C_2 = (4.6, 3.0)$ (***sol2***)

```
C = matrix(c(2.2, 1.6,
             4.6, 3.0), nrow = 2, ncol =2, byrow = TRUE)
df = df %>%
  mutate(d_C1 = sqrt((X1 - C[1, 1])^2 + (X2 - C[1, 2])^2),
         d_C2 = sqrt((X1 - C[2, 1])^2 + (X2 - C[2, 2])^2),
         Y = ifelse(d_C1 < d_C2,
                    1,
                    2))
J(df, C)
```

```
## [1] 9.521839
```

$J(y_{sol2}) = 9.521839$

Y para J^* (con distancias Euclídeas al cuadrado),

```
## función costo J^*(y)
J2 <- function(df, C) {
  J= df %>%
    group_by(Y) %>%
    summarise(d = ifelse(Y ==1,
                         (X1 - C[1, 1])^2 + (X2 - C[1, 2])^2,
                         (X1 - C[2, 1])^2 + (X2 - C[2, 2])^2)) %>%
    ungroup
  return(sum(J$d))
}
```

Para la solución $C_1 = (2.66667, 1.5)$ y $C_2 = (4.5000000, 3.5)$ (***sol1***):

```
C = matrix(c(2.666667, 1.5,
             4.500000, 3.5), nrow = 2, ncol =2, byrow = TRUE)
df = df %>%
  mutate(d_C1 = sqrt((X1 - C[1, 1])^2 + (X2 - C[1, 2])^2),
         d_C2 = sqrt((X1 - C[2, 1])^2 + (X2 - C[2, 2])^2),
         Y = ifelse(d_C1 < d_C2,
                    1,
                    2))
J2(df, C)
```

```
## [1] 12.83333
```

$J^*(y_{sol1}) = 12.83333$.

Para la solución $C_1 = (2.2, 1.6)$ y $C_2 = (4.6, 3.0)$ (***sol2***):

```
C = matrix(c(2.2, 1.6,
             4.6, 3.0), nrow = 2, ncol =2, byrow = TRUE)
df = df %>%
  mutate(d_C1 = sqrt((X1 - C[1, 1])^2 + (X2 - C[1, 2])^2),
         d_C2 = sqrt((X1 - C[2, 1])^2 + (X2 - C[2, 2])^2),
         Y = ifelse(d_C1 < d_C2,
                    1,
                    2))
J2(df, C)
```

```
## [1] 11.2
```

$J^*(y_{sol2}) = 11.2$

En ambos casos la solución segunda parece dar mejores resultados:

$$J(y_{sol1}) = 9.823299 > J(y_{sol2}) = 9.521839,$$
$$J^*(y_{sol1}) = 12.83333 > J^*(y_{sol2}) = 11.2.$$

6.5.4) K-means se puede ejecutar de forma automática en R

El algoritmo K-means se puede ejecutar de forma automática en R con el comando Kmeans.

Por defecto, usa el algoritmo de Hartigan and Wong (1979).

Para ejecutarlo en este ejemplo:

```
X1 <- c(1, 2, 3, 2, 5, 5, 3, 4, 4, 5)
X2 <- c(2, 1, 1, 2, 1, 3, 2, 3, 4, 4)
d <- data.frame(X1, X2)
CA <- kmeans(d, 2)
CA$centers
```

```
##    X1  X2
## 1 4.6 3.0
## 2 2.2 1.6
```

Los grupos coinciden con los obtenidos en la segunda solución anterior (óptima).

También se pueden guardar

```
CA1 <- CA$centers[1,]
CA2 <- CA$centers[2,]
```

Los grupos se obtienen con

```
CA$cluster
```

```
## [1] 2 2 2 1 1 2 1 1 1
```

La solución coincide con la representada en la gráfica.

Las sumas de las distancias al cuadrado en los grupos se obtienen con

```r
CA$withinss
```

```
## [1] 7.2 4.0
```

obteniendo $7.2 + 4.0 = 11.2$ (como antes)

El comando

```r
CA$totss
```

```
## [1] 30.5
```

proporciona la suma de las distancias al cuadrado sin grupos (o con un único grupo) obteniéndose 30.35.

Si se calcula directamente,

```r
centroides = df %>%
  summarise(C_x1 = mean(X1),
            C_x2 = mean(X2))
C[1, ] = c(centroides$C_x1[1], centroides$C_x2[1])
df = df %>%
  mutate(d = (X1 - C[1, 1])^2 + (X2 - C[1, 2])^2)
sum(df$d)
```

```
## [1] 30.5
```

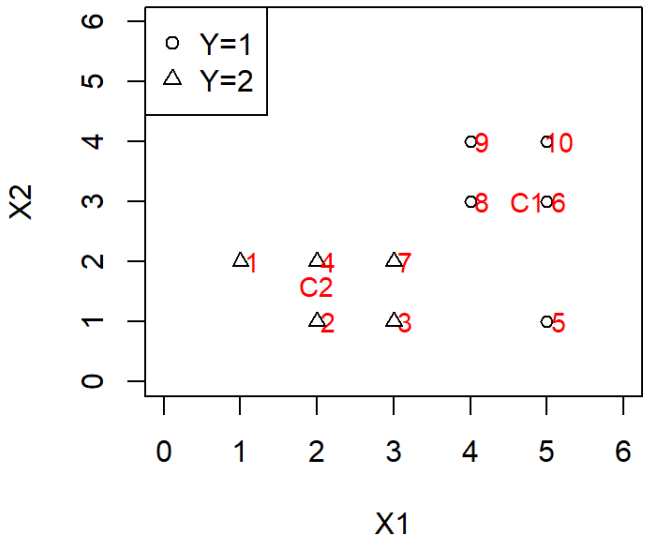
Por lo que al agruparlos se ha producido una disminución del

$$1 - \frac{11.2}{30.5} = 0.6327869$$

por uno, es decir, con dos grupos la variabilidad se reduce un 63.28%.

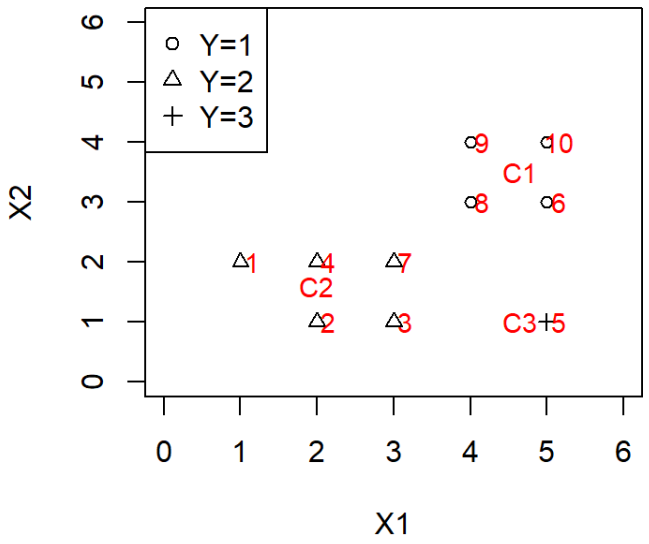
Individuos agrupados con Kmeans en 2 grupos:

```r
K <- 2
CA <- kmeans(d, K)
plot(X1, X2, xlab = "X1", ylab = "X2", pch = as.integer(CA$cluster),
     xlim = c(0, 6), ylim = c(0, 6), cex = 0.9)
legend('topleft', legend = c('Y=1', 'Y=2'), pch = 1:K, cex = 1)
text(CA$centers[1, 1] + 0.15, CA$centers[1, 2], 'C1', cex = 0.9, col = 'red')
text(CA$centers[2, 1] - 0.20, CA$centers[2, 2], 'C2', cex = 0.9, col = 'red')
text(X1 + 0.15, X2, 1:length(X1), cex = 0.9, col = 'red')
```



Individuos agrupados con `Kmeans` en 3 grupos (hay un grupo que tiene un único dato, por lo tanto, será su centroide):

```
K <- 3
CA <- kmeans(d, K)
plot(X1, X2, xlab = "X1", ylab = "X2", pch = as.integer(CA$cluster),
     xlim = c(0, 6), ylim = c(0, 6), cex = 0.9)
legend('topleft', legend = c('Y=1', 'Y=2', 'Y=3'), pch = 1:K, cex = 1)
text(CA$centers[1, 1] + 0.15, CA$centers[1, 2], 'C1', cex = 0.9, col = 'red')
text(CA$centers[2, 1] - 0.20, CA$centers[2, 2], 'C2', cex = 0.9, col = 'red')
text(CA$centers[3, 1] - 0.35, CA$centers[3, 2], 'C3', cex = 0.9, col = 'red')
text(X1 + 0.15, X2, 1:length(X1), cex = 0.9, col = 'red')
```



Al ejecutar este comando pueden aparecer otras soluciones porque las soluciones dependen de los valores iniciales.

La función `kmeans` permite ejecutar el algoritmo con diferentes valores de partida (argumento `nstart`) y proporcionar la mejor de esas

soluciones.

```
CA <- kmeans(d, 3, nstart = 10)
```

Con tres grupos la variabilidad se reduce en un 80.3 %.

```
CA = kmeans(d, 3, nstart = 10)
1- (sum(CA$withinss)/CA$totss)
```

```
## [1] 0.9032787
```

6.6) Método jerárquico

6.6.1) Índice de similaridad

En este caso no fijamos de antemano un número de grupos.

Lo que haremos es, dada una instancia, definir un índice de similaridad entre dos observaciones con

$$I(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = 1 - \frac{d(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})}{\max\limits_{r,s} d(\mathbf{x}^{(r)}, \mathbf{x}^{(s)})} \in [0, 1].$$

Se puede dar una definición análoga para grupos.

La idea es establecer clasificaciones calculando los índices de similitud (o distancias) que se van obteniendo.

Finalmente, dependiendo del índice de similitud elegido, obtendremos un número determinado de grupos (uniendo los que tienen similitud menor que ese índice).

6.6.2) Algoritmos

Consideraremos dos algoritmos.

En el primero partiremos de n grupos formados por un individuo cada uno.

- En el primer paso uniremos las dos observaciones más cercanas (distancia mínima) que serán las que tengan un índice de similitud mayor.
- Recalculamos las distancias para estos grupos y unimos los dos grupos más cercanos, continuamos así hasta conseguir un único grupo.

En el segundo, procederemos de forma inversa, partiremos de un único grupo formado por todas las observaciones.

- En un primer paso separaremos en dos de forma que las distancias entre estos dos grupos sea máxima (o las distancias a esos dos grupos de sus individuos sea mínima).
- En el siguiente paso formaremos un tercer grupo tomando individuos de los grupos 1 y 2 con un criterio similar.
- Procederemos así hasta conseguir n grupos.

Claramente, este método es más lento que el anterior.

6.6.3) Un ejemplo sencillo

Para ver un ejemplo analizaremos los datos del ejemplo anterior usando el primer método.

En primer lugar calculamos las distancias Euclídeas entre todos los individuos:

```
n <- length(X1)
D <- matrix(NA, n, n)
for (i in 1:n) {
  for (j in 1:n) {
    D[i, j] <- dE(d[i, ], d[j, ])
    }
}
```

D	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}
O_1	0	1.41	2.24	**1**	4.12	4.12	2	3.16	3.61	**4.47**
O_2	1.41	0	1	1	3	3.61	1.41	2.83	3.61	4.24
O_3	2.24	1	0	1.41	2	2.83	1	2.24	3.16	3.61
O_4	1	1	4.41	0	3.16	3.16	1	2.24	2.83	3.61
O_5	4.12	3	2	3.16	0	2	2.24	2.24	3.16	3
O_6	4.12	3.61	2.83	3.16	2	0	2.24	1	1.41	1
O_7	2	1.41	1	1	2.24	2.24	0	1.41	2.24	2.83
O_8	3.16	2.83	2.24	2.24	1	1	1.41	0	1	1.41
O_9	3.61	3.61	3.16	2.83	1.41	1.41	2.24	1	0	1
O_{10}	**4.47**	4.24	3.61	3.61	1	1	1.41	1.41	1	0

Observamos que el máximo se alcanza en $D_{1,10} = 4.47$ y el mínimo eliminando los ceros es 1 y se alcanza en varios puntos (esto se debe a que los puntos son discretos).

El primero que detecta el programa es $D_{1,4} = 1$ por lo que será el primer grupo $G_1 = \{1, 4\}$.

El índice de similaridad será

$$I(\mathbf{x}^{(1)}, \mathbf{x}^{(4)}) = 1 - \frac{d(\mathbf{x}^{(1)}, \mathbf{x}^{(4)})}{\max\limits_{r,s} d(\mathbf{x}^{(r)}, \mathbf{x}^{(s)})} = 1 - \frac{1}{4.472136} = 0.7763932.$$

El siguiente paso dependerá de la distancia entre grupos elegida.

Si queremos mantener esas distancias y detectar esos empates debemos elegir la distancia del vecino más próximo (D_2).

Todas las demás nos darán valores mayores. Con esta distancia (tras varias iteraciones) uniríamos todos los puntos que están a distancia 1 de alguno del grupo obteniendo:

$$G_1 = \{1, 2, 3, 4, 7\},\ G_2 = \{5\},\ G_3 = \{6, 8, 9, 10\}.$$

La matriz de distancias D_2 para estos tres grupos será

D_2	G_1	G_2	G_3
G_1	0	2	**1.41**
G_2	2	0	2
G_3	**1.41**	2	0

El mínimo se alcanza en $d(G_1, G_3) = d(\mathbf{x}^{(7)}, \mathbf{x}^{(8)}) = 1.41$.

Por lo que en el segundo paso uniríamos los grupos G_1 y G_3 que tendrán un índice de similaridad

$$I(\mathbf{x}^{(7)}, \mathbf{x}^{(8)}) = 1 - \frac{d(\mathbf{x}^{(7)}, \mathbf{x}^{(8)})}{\max\limits_{r,s} d(\mathbf{x}^{(r)}, \mathbf{x}^{(s)})} = 1 - \frac{1.41}{4.472136} = 0.6847144.$$

En el último paso uniríamos el grupo G_2 con $G_1 \cup G_3$ a distancia 2 y similaridad

$$I(\mathbf{x}^{(3)}, \mathbf{x}^{(5)}) = 1 - \frac{d(\mathbf{x}^{(3)}, \mathbf{x}^{(5)})}{\max\limits_{r,s} d(\mathbf{x}^{(r)}, \mathbf{x}^{(s)})} = 1 - \frac{2}{4.472136} = 0.5527864.$$

El dendograma debe mostrar estas uniones usando las distancias o los índices de similaridad.

Con otras distancias y/o usando el segundo método (que parte de un único grupo que se separa en dos) podemos obtener resultados diferentes.

También observamos que los resultados no tienen por qué coincidir con los obtenidos con el algoritmo K-medias.

La elección de un método u otro dependerá de los datos que tengamos y del problema que se quiera resolver (costo).

- Por ejemplo, si lo que queremos es agrupar a los usuarios para ser atendidos por centros deberemos usar distancias basadas en centroides que representarán dónde se situarán (aproximadamente) esos centros.
- Por contra, si lo que queremos es simplemente clasificar empresas o países según sus características, estos centroides no serán tan importantes.

6.6.4) ¿Cómo realizar este agrupamiento de forma automática en R?

Podemos usar la función `hclust`.

En primer lugar calcularemos las distancias con

```
D <- dist(d, method = 'euclidean')
```

representadas en forma de vector.

Para visualizarlas en forma de matriz usaremos `as.matrix(D)[1:10, 1:10]`.

Ahora, para hacer un CA utilizamos la instrucción

```
CA2 <- hclust(D, method = 'complete')
```

Hemos usado el método de agrupación `complete` que usa la distancia del vecino más lejano (es el que usa `R` por defecto).

Otras opciones son:

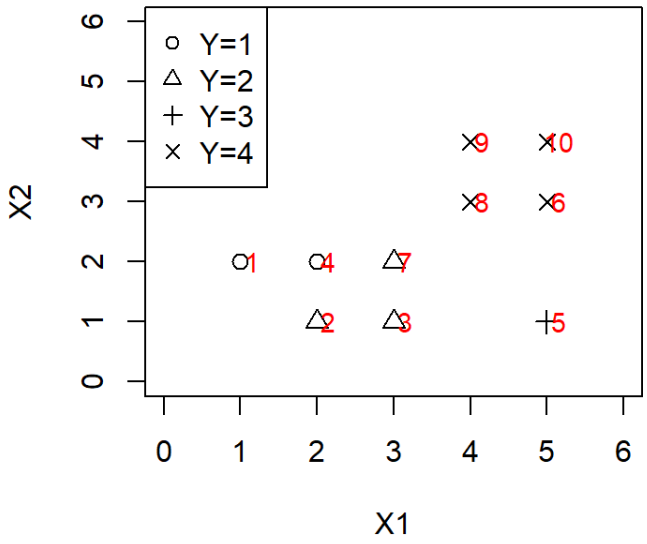
- `single:` vecino más cercano
- `average:` media de todas las distancias entre todas las parejas de puntos de los dos grupos
- `centroid:` distancia entre los centroides.

Para $K = 4$ grupos

```
grupos <- cutree(CA2, k = 4)
```

obtendremos los grupos:

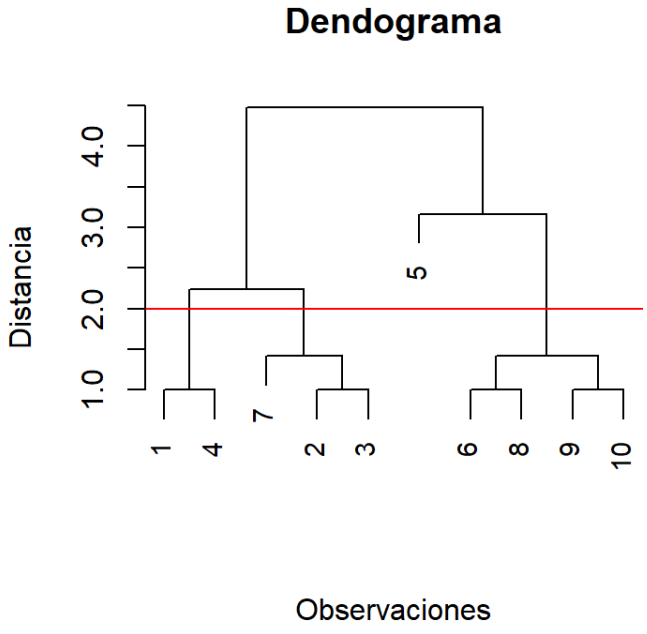
```
plot(X1, X2, pch = as.integer(grupos),
     xlim = c(0, 6), ylim = c(0, 6), cex = 1.2)
legend('topleft', legend = c('Y=1', 'Y=2', 'Y=3', 'Y=4'),
       pch = 1:4, cex = 1)
text(X1 + 0.15, X2, 1:n, cex = 0.9, col = 'red')
```



Representación del dendograma.

La línea roja en el dendograma representa la distancia que nos da 4 grupos.

```
plot(CA2, cex = 0.9, main = 'Dendograma',
     ylab = 'Distancia', xlab = 'Observaciones', sub = '')
abline(h = 2, col = 'red')
```



En el dendograma primero se unen los puntos que están a menor distancia (en este caso era distancia 1) y, posteriormente se calculan las distancias entre grupos usando la distancia al vecino más lejano (D_3).

Por ejemplo, la distancia de la observación 7 al grupo $\{2, 3\}$ es

```
dE(d[7, ], d[2, ])
```

```
## [1] 1.414214
```

es decir, 1.414214.

Lo mismo ocurre con la distancia entre los grupos $\{6, 8\}$ y $\{9, 10\}$. La distancia mayor es la de observación 5 al grupo $\{6, 8, 9, 10\}$ obtenida con

```
dE(d[5, ], d[9, ])
```

```
## [1] 3.162278
```

y que vale 3.162278.