Universidad Nacional de Córdoba

Facultad de Matemática, Astronomía, Física y Computación

Trabajo Especial Licenciatura en Matemática

Métodos clásicos de clasificación: comparación y aplicación.

Facundo Eduardo Godoy
Directora: Aldana González Montoro

Julio del 2021

Índice general

Capítulo 1

Introducción

En este trabajo estudiamos métodos clásicos de clasificación supervisada. Abordaremos la teoría basándonos en Peña (2002) y James et al. (2013). Para cada método estudiado se resuelven los ejercicios propuestos en Peña (2002), algunos de forma manual y otros utilizando el software estadístico R Core Team (2020). En este último caso se muestra el código utilizado. Realizamos también un estudio de simulación para comparar los métodos estudiados en distintos escenarios. Estas simulaciones son presentadas en el capítulo 7. En el capítulo 8 presentamos una aplicación a datos reales en el contexto de predicción de la distancia de dos celulares inteligentes usando señales de Bluetooth.

La mayoría de los problemas de clasificación se dividen en una de dos categorías: supervisadas o no supervisadas. La diferencia es que, en la clasificación supervisada, para cada observación, tenemos la medición del predictor $\mathbf{x}_i$, $i = 1, \ldots, n$ y una medida de respuesta asociada y_i. Deseamos ajustar un modelo que relacione la respuesta con los predictores, con el objetivo de predecir con precisión la respuesta para futuras observaciones (predicción) o comprender mejor la relación entre la respuesta y los predictores (inferencia).

Por el contrario, la clasificación no supervisada describe la situación algo más desafiante en la que para cada observación $i = 1, \ldots, n$, observamos un vector de medidas $\mathbf{x}_i$ pero ninguna respuesta asociada y_i. En este entorno, en cierto sentido estamos trabajando a ciegas; la situación se denomina no supervisada porque carecemos de una variable de respuesta que pueda supervisar nuestro análisis.

El problema de discriminación o clasificación supervisada, que abordamos en este trabajo, puede plantearse de varias formas y aparece en muchas áreas de la actividad humana. El planteamiento estadístico del problema es el siguiente. Se dispone de un conjunto amplio de elementos que pueden venir de dos o más poblaciones distintas. En cada elemento se ha observado una variable aleatoria $p-$ dimensional $\mathbf{x}$, a la que en la literatura se le suele llamar predictoras,

covariables o características. Se desea clasificar un nuevo elemento, con valores de las variables conocidas, en una de las poblaciones.

El problema de discriminación aparece en muchas situaciones en que necesitamos clasificar elementos con información incompleta. Por ejemplo, los sistemas automáticos de concesión de créditos (credit scoring) implantados en muchas instituciones financieras tienen que utilizar variables medibles hoy (ingresos, antigüedad en el trabajo, patrimonio, etc) para prever el comportamiento futuro. En otros casos la información podría estar disponible, pero puede requerir destruir el elemento, como en el control de calidad de la resistencia a la tensión de unos componentes. Finalmente, en otros casos la información puede ser muy costosa de adquirir. En ingeniería este problema se ha estudiado con el nombre de reconocimiento de patrones (pattern recognition), para diseñar máquinas capaces de clasificar de manera automática. Por ejemplo, reconocer voces y sonidos, clasificar billetes o monedas, reconocer caracteres escritos en una pantalla de ordenador o clasificar cartas según el distrito postal.

1.1 ¿Por qué tantos métodos?

En este trabajo estudiamos la teoría y la práctica de estos 4 métodos clásicos de clasificación: Análisis Discriminante Lineal (LDA), Análisis Discriminante Cuadrático (QDA), Regresión Logistica y K vecinos más próximos (K-NN).

¿Por qué es necesario introducir tantos enfoques de clasificación diferentes, en lugar de solo uno que sea mejor a todos? Lo que trataremos de ver en este trabajo, es que ningún método domina a todos los demás en todos los conjuntos de datos posibles. En un conjunto de datos en particular, un método específico puede funcionar mejor, pero algún otro método puede funcionar mejor en un conjunto de datos similar pero diferente. Por lo tanto, es una tarea importante decidir para cualquier conjunto de datos determinado qué método produce los mejores resultados. Seleccionar el mejor enfoque puede ser una de las partes más desafiantes de realizar la clasificación en la práctica.

1.2 El clasificador de Bayes

Es posible demostrar (lo haremos en la sección siguiente para el caso de dos poblaciones) que la tasa de error de prueba dada en (2.1) se minimiza, en promedio, por un clasificador muy simple que asigna cada observación a la clase más probable, dado el valor sus predictoras. En otras palabras, simplemente deberíamos asignar una observación de prueba con el vector predictor $\mathbf{x}_0$ a la clase j para la cual

$$\mathrm{P}\left(Y = j \mid X = \mathbf{x}_0\right) \tag{1.1}$$

es más grande. Tenga en cuenta que (1.1) es una probabilidad condicional: es la probabilidad de que $Y = j$, dado el vector predictor observado $\mathbf{x}_0$. Este clasificador muy simple se llama clasificador de Bayes.

El clasificador de Bayes produce la tasa de error de prueba más baja posible, denominada tasa de error de Bayes. Dado que el clasificador de Bayes siempre elegirá la clase para la cual (1.1) es más grande, la tasa de error en $X = \mathbf{x}_0$ será $1 - \max_j \mathrm{P}(Y = j \mid X = \mathbf{x}_0)$. En general, la tasa de error general de Bayes viene dada por

$$1 - E\left(\max_j \mathrm{P}(Y = j \mid X)\right)$$

donde la expectativa promedia la probabilidad sobre todos los valores posibles de X. La tasa de error de Bayes es análoga al error irreducible.

1.3 Clasificación de Bayes entre dos poblaciones

1.3.1 Planteamiento del Problema

Sean P_1 y P_2 dos poblaciones donde tenemos definida una variable aleatoria vectorial, $\mathbf{X}$, p-variante. Supondremos que $\mathbf{X}$ es absolutamente continua y que las funciones de densidad de ambas poblaciones, f_1 y f_2, son conocidas. Vamos a estudiar el problema de clasificar un nuevo elemento, $\mathbf{x}_0$, con valores conocidos de las p variables en una de estas poblaciones. Si conocemos las probabilidades a priori π_1,π_2, con $\pi_1 + \pi_2 = 1$, de que el elemento venga de cada una de las dos poblaciones, su distribución de probabilidad será una distribución mezclada

$$f(\mathbf{x}) = \pi_1 f_1(\mathbf{x}) + \pi_2 f_2(\mathbf{x})$$

y una vez observado $\mathbf{x}_0$ podemos calcular las probabilidades a posteriori de que el elemento haya sido generado por cada una de las dos poblaciones, $P(i|\mathbf{x}_0)$, con $i = 1, 2$. Estas probabilidades se calculan por el teorema de Bayes

$$P(1|\mathbf{x}_0) = \frac{P(\mathbf{x}_0|1)\pi_1}{\pi_1 P(\mathbf{x}_0|1) + \pi_2 P(\mathbf{x}_0|2)}$$

(La notación $P(1|\mathbf{x}_0)$ significa $P(Y = 1|X = \mathbf{x}_0)$)

y como $P(\mathbf{x}_0|1) = f_1(\mathbf{x}_0)\Delta\mathbf{x}_0$, tenemos que:

$$P(1|\mathbf{x}_0) = \frac{f_1(\mathbf{x}_0)\pi_1}{f_1(\mathbf{x}_0)\pi_1 + f_2(\mathbf{x}_0)\pi_2} \tag{1.2}$$

y para la segunda población

$$P(2|\mathbf{x}_0) = \frac{f_2(\mathbf{x}_0)\pi_2}{f_1(\mathbf{x}_0)\pi_1 + f_2(\mathbf{x}_0)\pi_2} \tag{1.3}$$

Clasificaremos $\mathbf{x}_0$ en la población más probable a posteriori. Como los denominadores son iguales, clasificaremos $\mathbf{x}_0$ en P_2 si:

$$\pi_2 f_2(\mathbf{x}_0) > \pi_1 f_1(\mathbf{x}_0)$$

Si las probabilidades a priori son iguales, la condición de clasificar en P_2 se reduce a:

$$f_2(\mathbf{x}_0) > f_1(\mathbf{x}_0)$$

es decir, clasificamos a $\mathbf{x}_0$ en la población más probable, o donde su verosimilitud es más alta.

1.3.2 Consideración de las consecuencias

En muchos problemas de clasificación los errores que podemos cometer tienen distintas consecuencias que podemos cuantificar. Por ejemplo, si una máquina automática clasifica equivocadamente un billete de 10 euros como de 20, y devuelve el cambio equivocado, el coste de clasificación es de 10 euros. En otros casos estimar el coste puede ser más complejo: si no concedemos un crédito que sería devuelto podemos perder un cliente y los ingresos futuros que este podría generar, mientras que si el crédito no se devuelve el coste es la cantidad impagada. Como tercer ejemplo, si clasificamos un proceso productivo como en estado de control, el coste de equivocarnos será una producción defectuosa, y si, por error, paramos un proceso que funciona adecuadamente, el coste será el de la parada y revisión.

En general supondremos que las posibles decisiones en el problema son únicamente dos: asignar en P_1 o en P_2. Una regla de decisión es una partición del espacio muestral E_x (que en general será $\mathbb{R}^p$) en dos regiones A_1 y $A_2 = E_x - A_1$, tales que:

$$\text{si } \mathbf{x}_0 \in A_1 \Longrightarrow d_1 \text{ (clasificar en } P_1).$$
$$\text{si } \mathbf{x}_0 \in A_2 \Longrightarrow d_2 \text{ (clasificar en } P_2).$$

Si las consecuencias de un error de clasificación pueden cuantificarse, podemos incluirlas en la solución del problema formulándolo como un problema bayesiano de decisión. Supongamos que:

1. las consecuencias asociadas a los errores de clasificación son, $c(2|1)$ y $c(1|2)$, donde $c(i|j)$ es el coste de clasificación en P_i de una unidad que pertenece a P_j. Estos costes se suponen conocidos;

2. el decisor quiere maximizar su función de utilidad y esto equivale a minimizar el coste esperado.

Con estas dos hipótesis la mejor decisión es la que minimiza los costes esperados, o funciones de pérdida de oportunidad, en la terminología de Wald. Los resultados de cada decisión que se presenta esquemáticamente en la Figura 1.1 Si clasificamos al elemento en el grupo 2 las posibles consecuencias son:

(a) acertar, con probabilidad $P(2|\mathbf{x}_0)$, en cuyo caso no hay ningún coste de penalización;
(b) equivocarnos, con probabilidad $P(1|\mathbf{x}_0)$, en cuyo caso incurrimos en el coste asociado $c(2|1)$.

El coste promedio, o valor esperado, de la decisión " d_2: clasificar $\mathbf{x}_0$ en P_2" será:

$$E(d_2) = c(2|1)P(1|\mathbf{x}_0) + 0P(2|\mathbf{x}_0) = c(2|1)P(1|\mathbf{x}_0).$$

Análogamente, el coste esperado de la decisión " d_1: clasificar $\mathbf{x}_0$ en el grupo 1" es:

$$E(d_1) = 0P(1|\mathbf{x}_0) + c(1|2)P(2|\mathbf{x}_0) = c(1|2)P(2|\mathbf{x}_0).$$

Asignaremos al $\mathbf{x}_0$ elemento al grupo 2 si su coste esperado es menor, es decir, utilizando (1.2) y (1.3), si:

$$\frac{f_2(\mathbf{x}_0)\pi_2}{c(2|1)} > \frac{f_1(\mathbf{x}_0)\pi_1}{c(1|2)}$$

Esta condición indica que, a igualdad de los otros términos, clasificaremos en la población P_2 si:

(a) su probabilidad a priori es más alta;

(b) la verosimilitud de que $\mathbf{x}_0$ provenga de P_2 es más alta;

(c) el coste de equivocarnos al clasificarlo en P_2 es más bajo.

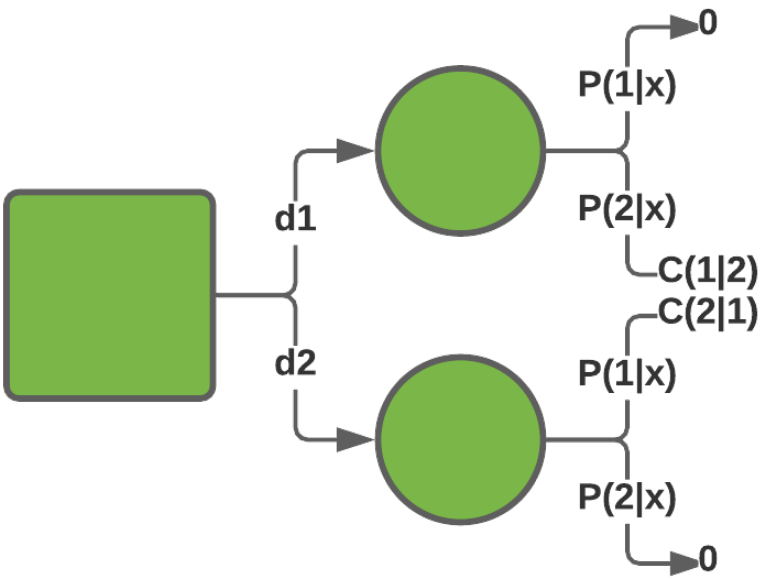


Figura 1.1: Representación de un problema de clasificación entre dos grupos como un problema de decisión.

1.3.3 Pueba de que la tasa de error de Bayes es óptima

El criterio de minimizar la probabilidad de error puede escribirse como minimizar P_T , donde:

$$P_T(\text{error}) = P(1 \mid \mathbf{x} \in 2) + P(2 \mid \mathbf{x} \in 1)$$

siendo $P(i|x \in j)$ la probabilidad de clasificar en la población i una observación que proviene de la j. Esta probabilidad viene dada por el área encerrada por la distribución j en la zona de clasificación de i, es decir:

$$P(i \mid \mathbf{x} \in j) = \int_{A_i} f_j(\mathbf{x}) d\mathbf{x}$$

por tanto:

$$P_T = \int_{A_1} f_2(\mathbf{x}) d\mathbf{x} + \int_{A_2} f_1(\mathbf{x}) d\mathbf{x}$$

y como A_1 y A_2 son complementarios:

$$\int_{A_1} f_2(\mathbf{x}) d\mathbf{x} = \mathbf{1} - \int_{A_2} f_2(\mathbf{x}) d\mathbf{x}$$

que conduce a:

$$P_T = 1 - \int_{A_2} \left(f_2(\mathbf{x}) - f_1(\mathbf{x})\right) d\mathbf{x}$$

y para minimizar la probabilidad de error debemos maximizar la integral. Esto se consigue definiendo A_2 como el conjunto de puntos donde el integrando es positivo, es decir:

$$A_2 = \{\mathbf{x} \mid f_2(\mathbf{x}) > f_1(\mathbf{x})\}$$

y obtenemos de nuevo el criterio antes establecido.

Capítulo 2

Preliminares

2.1 ¿Cómo comparar los métodos?

Para evaluar el rendimiento de un método de clasificación en un conjunto de datos dado, necesitamos alguna forma de medir qué tan bien sus predicciones coinciden realmente con los datos observados. Es decir, necesitamos cuantificar la medida en que el valor de respuesta predicho para una observación determinada se acerca al valor de respuesta real para esa observación.

El enfoque más común para cuantificar la precisión de nuestro clasificador es la tasa de error de entrenamiento (ó de prueba), la proporción de errores que se cometen si aplicamos nuestro clasificador a un conjunto de observaciones cuya clasificación conocemos.

$$\frac{1}{n}\sum_{i=1}^{n} I\left(y_i \neq \widehat{y}_i\right) \tag{2.1}$$

Aquí $\widehat{y}_i$ es la etiqueta de clase predicha para la $i-$ésima observación usando el clasificador (o función de clasificación). Donde $I\left(y_i \neq \widehat{y}_i\right)$ es una variable indicadora que es igual a 1 si $y_i \neq \widehat{y}_i$ y cero si $y_i = \widehat{y}_i$. Si $I\left(y_i \neq \widehat{y}_i\right) = 0$ entonces la $i-$ésima observación fue clasificada correctamente por nuestro método de clasificación; de lo contrario, se clasificó erróneamente. Por tanto, la ecuación (2.1) calcula la fracción de clasificaciones incorrectas.

2.2 Muestra de entrenamiento y de prueba

Para no subestimar la tasa de error de los modelos las muestras suelen separasre en dos subconjuntos: **datos de enrtenamiento o muestra de**

entrenamiento y **datos de prueba**. Los **datos de enrtenamiento** son aquellos que utilizamos para ajustar o entrenar nuestro clasificador. Y los **datos de prueba** los que utilizamos para evaluar nuestro clasificador, son datos que conocemos su clase o población, pero no se usan para ajustar al modelo, se usan para calcular la tasa de error.

2.3 La matriz de confusión y sus métricas

Matriz de confusión: Es una herramienta que permite visualizar el desempeño de un algoritmo de clasificación supervisada. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real, o sea en términos prácticos nos permite ver qué tipos de aciertos y errores está teniendo nuestro modelo a la hora de ajustarse a los datos observados. En una matriz de confusión, dónde la variable de respuesta es binaria, tenemos 4 opciones **Verdadero Positivo (VP)**, **Verdadero Negativo (VN)**, **Falso Positivo (FP)**, **Falso Negativo (FN)** (Esto no tiene sentido cuando hay 3 o más clases ya que no hay noción de positivo o negativo, pero el razonamiento que veremos a continuación es análogo para esos casos).

VP: es la cantidad de positivos que fueron clasificados correctamente como positivos por el modelo.
VN: es la cantidad de negativos que fueron clasificados correctamente como negativos por el modelo.
FN: es la cantidad de positivos que fueron clasificados incorrectamente como negativos.
FP: es la cantidad de negativos que fueron clasificados incorrectamente como positivos.

Exactitud (Acurracy): Porcentaje de los datos clasificados correctamente

$$\text{Exactitud } = \frac{VP + VN}{\text{Total}}$$

Precisión : Es la proporción entre el número de predicciones correctas (tanto positivas como negativas) y el total de predicciones.

$$\text{Precisión } = \frac{VP}{\text{Total clasificados positivos}}$$

Tasa de error: Porcentaje de los datos clasificados incorrectamente

$$\text{Tasa de error } = \frac{FP + FN}{\text{Total}}$$

Sensibilidad : Es la proporción de casos positivos que fueron correctamente identificadas por el algoritmo.

$$\text{Sensibilidad } = \frac{VP}{\text{Total Positivos}}$$

Especificidad : Es la proporción de casos negativos que fueron correctamente identificadas por el algoritmo.

$$\text{Especificidad } = \frac{VN}{\text{Total Negativos}}$$

2.4 Validación cruzada

El conjunto de validación es una estrategia para estimar el error de prueba asociado con el ajuste de un método de clasificación particular en un conjunto de observaciones. Implica dividir aleatoriamente el conjunto de observaciones disponibles en dos partes, un conjunto de entrenamiento y un conjunto de prueba. El modelo se ajusta al conjunto de entrenamiento y el modelo ajustado se utiliza para predecir las respuestas de las observaciones en el conjunto de prueba. La tasa de error del conjunto de prueba resultante, que generalmente se evalúa mediante (2.1) , proporciona una estimación de la tasa de error de la prueba real.

Una forma de dividir el conjunto de observaciones, y además la que usaremos en este trabajo, es Leave-one-out cross-validation (LOOCV) implica dividir el conjunto de observaciones en dos partes, una única observación (x_1, y_1) para el conjunto de validación, y las observaciones restantes $(x_2, y_2), \ldots, (x_n, y_n)$ forman el conjunto de entrenamiento. El método de clasificación se ajusta a las $n-1$ observaciones de entrenamiento, y se hace una predicción $\widehat{y}_1$ para la observación excluida, usando su valor x_1. Y llamaremos $Error_1 = I\left(y_1 \neq \widehat{y}_1\right)$. Podemos repetir el procedimiento seleccionando (x_2, y_2) para los datos de prueba, entrenando el procedimiento de aprendizaje estadístico en las $n-1$ observaciones $(x_1, y_1), (x_3, y_3), \ldots, (x_n, y_n)$, y calculando $Error_2 = I\left(y_2 \neq \widehat{y}_2\right)$. Repetir este enfoque n veces produce n errores, $Error_1, \ldots, Error_n$. La estimación de LOOCV para la tasa de error es el promedio de estas n estimaciones de errores de prueba:

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \text{Error}_i$$

Una alternativa a LOOCV es k-fold CV. Este enfoque implica dividir aleatoriamente el conjunto de observaciones en k grupos, de aproximadamente el mismo tamaño. El primer grupo se trata como un conjunto de prueba y el

método se ajusta a los $k-1$ grupos restantes. La tasa de error, $Error_1$, se calcula luego sobre las observaciones en el grupo retenido. Este procedimiento se repite k veces; cada vez, un grupo diferente de observaciones se trata como un conjunto de prueba. Este proceso da como resultado k estimaciones de la tasa de error de prueba, $Error_1, Error_2, \ldots, Error_k$. La estimación de CV de k veces se calcula promediando estos valores,

$$\mathrm{CV}_{(k)} = \frac{1}{k} \sum_{i=1}^{k} \mathrm{Error}_i$$

En este trabajo utilizaremos esta estrategia no solo para estimar la tasa de error de prueba real, también la utilizaremos para optimizar parametros de los modelos, por ejemplo la elección del K para K-NN que veremos en la sección 6.4.1.

Capítulo 3

Análisis Discriminante Lineal (LDA)

3.1 Idea intuitiva

El Análisis Discriminante Lineal o Linear Discrimiant Analysis (LDA) es un método de clasificación supervisado para variables cualitativas. En el que un conjunto de observaciones de dos o más grupos son conocidos a priori y nuevas observaciones se clasifican en uno de ellos en función de sus características. Haciendo uso del teorema de Bayes, LDA estima la probabilidad de que una observación, dado un determinado valor de los predictores, pertenezca a cada una de las clases de la variable cualitativa, $P(Y = k|X = \mathbf{x})$. Al igual que el clasificador de Bayes, asigna la observación a la clase k para la que la probabilidad predicha es mayor. Y la diferencia con LDA la veremos en la siguiente sección 3.2.

3.2 Poblaciones Normales: Función lineal discriminante

Vamos a aplicar el análisis visto en la sección 1.3 al caso en que f_1 y f_2 son distribuciones normales con distintos vectores de medias pero idéntica matriz de varianzas. Para establecer la regla con carácter general supondremos que se desea clasificar un elemento genérico $\mathbf{x}$, que si pertenece a la población $i = 1, 2$ tiene función de densidad:

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|V|^{1/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{x} - \mu_i\right)' \mathbf{V}^{-1}\left(\mathbf{x} - \mu_i\right)\right\}$$

La partición óptima, es, de acuerdo con la sección 1.3, clasificar en la población P_2 si:

$$\frac{f_2(\mathbf{x})\pi_2}{c(2|1)} > \frac{f_1(\mathbf{x})\pi_1}{c(1|2)}$$

Como ambos términos son siempre positivos, tomando logaritmos, sustituyendo $f_i(\mathbf{x})$ por su expresión y restando $\log\left(\frac{1}{(2\pi)^{p/2}|V|^{1/2}}\right)$ de ambos miembros de la desigualdad , la ecuación anterior se convierte en:

$$-\frac{1}{2}\left(\mathbf{x}-\mu_2\right)'\mathbf{V}^{-1}\left(\mathbf{x}-\mu_2\right)+\log\frac{\pi_2}{c(2|1)} > -\frac{1}{2}\left(\mathbf{x}-\mu_1\right)'\mathbf{V}^{-1}\left(\mathbf{x}-\mu_1\right)+\log\frac{\pi_1}{c(1|2)}$$

Llamando D_i^2 a la cantidad, denominada distancia de Mahalanobis entre el punto observado, $\mathbf{x}$, y la media de la población i:

$$D_i^2 = \left(\mathbf{x}-\mu_i\right)'\mathbf{V}^{-1}\left(\mathbf{x}-\mu_i\right)$$

podemos escribir:

$$-\frac{1}{2}D_2^2 + \log\frac{\pi_2}{c(2|1)} > -\frac{1}{2}D_1^2 + \log\frac{\pi_1}{c(1|2)} \tag{3.1}$$

y suponiendo iguales los costes y las probabilidades a priori, c(1|2) = c(2|1); π_1 = π_2, la regla anterior se reduce a:

$$\text{Clasificar en 2 si } D_1^2 > D_2^2$$

es decir, clasificar la observación en la población de cuya media esté más próxima, midiendo la distancia con la medida de Mahalanobis.

Observación : si las variables $\mathbf{x}$ tuvieran $\mathbf{V} = \mathbf{I}\,\sigma^2$, la regla equivale a utilizar la distancia euclídea.

Ejemplo 1: Se desea clasificar un retrato entre dos posibles pintores. Para ello se miden dos variables: la profundidad del trazo y la proporción que ocupa el retrato sobre la superficie del lienzo. Las medias de estas variables para el primer pintor, A, son (2 y 0.8) y para el segundo, B, (2.3 y 0.7) y las desviaciones típicas de estas variables son 0.5 y 0.1 y la correlación entre estas medidas es 0.5. La obra a clasificar tiene medidas de estas variables (2.1 y 0.75).

Las distancias de Mahalanobis serán, calculando la covarianza como el producto de la correlación por las desviaciones típicas:

$$D_A^2 = (2.1-2, 0.75-0.8) \begin{bmatrix} 0.25 & 0.025 \\ 0.025 & 0.01 \end{bmatrix}^{-1} \begin{pmatrix} 2.1-2 \\ 0.75-0.8 \end{pmatrix} = 0.52$$

y para la segunda

$$D_B^2 = (2.1-2.3, 0.75-0.7) \begin{bmatrix} 0.25 & 0.025 \\ 0.025 & 0.01 \end{bmatrix}^{-1} \begin{pmatrix} 2.1-2.3 \\ 0.75-0.7 \end{pmatrix} = 0.8133$$

Por tanto, asignaremos la obra al primer pintor.

3.3 Interpretación Geométrica

La regla general anterior puede escribirse de una forma equivalente que permite interpretar geométricamente el método de clasificación utilizado. La ecuación (3.1) indica que debemos calcular la distancia de Mahalanobis, corregirla por el término correspondiente a las probabilidades a priori y los costes, y clasificar en el población donde esta distancia modificada sea mínima.

$$\frac{1}{2}\left(\mathbf{x}-\mu_i\right)' \mathbf{V}^{-1}\left(\mathbf{x}-\mu_i\right) - \log\frac{\pi_i}{c(i|j)}$$

$$\frac{1}{2}\left(\mathbf{x}'\mathbf{V}^{-1}\mathbf{x} - \mu_i'\mathbf{V}^{-1}\mathbf{x} - \mathbf{x}'\mathbf{V}^{-1}\mu_i + \mu_i'\mathbf{V}^{-1}\mu_i\right) - \log\frac{\pi_i}{c(i|j)}$$

Recordemos que como $\mathbf{x}'\mathbf{V}^{-1}\mu_i \in \mathbb{R}$, entonces $\mathbf{x}'\mathbf{V}^{-1}\mu_i = \left(\mathbf{x}'\mathbf{V}^{-1}\mu_i\right)'$, y como $\mathbf{V}$ es simétrica, implica $\mathbf{V}^{-1} = (\mathbf{V}^{-1})'$, por lo tanto $\mathbf{x}'\mathbf{V}^{-1}\mu_i = \mu_i'\mathbf{V}^{-1}\mathbf{x}$

$$\frac{1}{2}\left(\mathbf{x}'\mathbf{V}^{-1}\mathbf{x} - 2\mu_i'\mathbf{V}^{-1}\mathbf{x} + \mu_i'\mathbf{V}^{-1}\mu_i\right) - \log\frac{\pi_i}{c(i|j)}$$

Como las distancias tiene siempre el término común $\mathbf{x}'\mathbf{V}^{-1}\mathbf{x}$, que no depende de la población, podemos eliminarlo de las comparaciones y calcular el indicador

$$-\mu_i'\mathbf{V}^{-1}\mathbf{x} + \frac{1}{2}\mu_i'\mathbf{V}^{-1}\mu_i - \log\frac{\pi_i}{c(i|j)}$$

que será una función lineal en $\mathbf{x}$ y clasificar el individuo en la población donde esta función sea mínima. Esta regla divide el conjunto de valores posibles de $\mathbf{x}$ en dos regiones cuya frontera viene dada por:

$$-\mu_1'\mathbf{V}^{-1}\mathbf{x} + \frac{1}{2}\mu_1'\mathbf{V}^{-1}\mu_1 = -\mu_2'\mathbf{V}^{-1}\mathbf{x} + \frac{1}{2}\mu_2'\mathbf{V}^{-1}\mu_2 - \log\frac{c(1|2)\pi_2}{c(2|1)\pi_1}$$

que, como función de $\mathbf{x}$, equivale a:

$$\left(\mu_2 - \mu_1\right)'\mathbf{V}^{-1}\mathbf{x} = \left(\mu_2 - \mu_1\right)'\mathbf{V}^{-1}\left(\frac{\mu_2 + \mu_1}{2}\right) - \log\frac{c(1|2)\pi_2}{c(2|1)\pi_1}$$

Llamando:

$$\mathbf{w} = \mathbf{V}^{-1}\left(\mu_2 - \mu_1\right) \tag{3.2}$$

la frontera puede escribirse como:

$$\mathbf{w}'\mathbf{x} = \mathbf{w}'\frac{\mu_2 + \mu_1}{2} - \log\frac{c(1|2)\pi_2}{c(2|1)\pi_1}$$

que es la ecuación de un hiperplano. En el caso particular en que c(1|2) π_2 = c(2|1) π_1, clasificaremos en P_2 si

$$\mathbf{w}'\mathbf{x} > \mathbf{w}'\left(\frac{\mu_1 + \mu_2}{2}\right)$$

o lo que es equivalente, si

$$\mathbf{w}'\mathbf{x} - \mathbf{w}'\mu_1 > \mathbf{w}'\mu_2 - \mathbf{w}'\mathbf{x} \tag{3.3}$$

Esta ecuación indica que el procedimiento para clasificar un elemento $\mathbf{x}_0$ puede resumirse como sigue:

(1) calcular el vector $\mathbf{w}$ con (3.2);

(2) construir la variable indicadora discriminante:

$$z = \mathbf{w}'\mathbf{x} = w_1x_1 + \ldots. + w_px_p$$

que transforma la variable multivariante $\mathbf{x}$ en la variable escalar z , que es una combinación lineal de los valores de la variable multivariante con coeficientes dados por el vector $\mathbf{w}$;

(3) calcular el valor de la variable indicadora para el individuo a clasificar, $\mathbf{x}_0 = (x_{10}, ..., x_{p0})$, con $z_0 = \mathbf{w}'\mathbf{x}_0$ y el valor de la variable indicadora para las medias de las poblaciones, $m_i = \mathbf{w}'\mu_i$. Clasificar en aquella población donde la distancia $|z_0 - m_i|$ sea mínima.

En términos de la variable escalar z, como el valor promedio de z en P_i es :

$$\mathbf{E}\left(z|P_i\right) = m_i = \mathbf{w}'\mu_i, \quad i = 1, 2$$

La regla de decisión (3.3) equivale a clasificar en P_2 si:

$$|z - m_1| > |z - m_2|$$

Esta variable indicadora, z, tiene varianza:

$$\mathrm{Var}(z) = \mathbf{w}'\,\mathrm{Var}(\mathbf{x})\mathbf{w} = \mathbf{w}'\mathbf{V}\mathbf{w} = \left(\mu_2 - \mu_1\right)'\mathbf{V}^{-1}\left(\mu_2 - \mu_1\right) = D^2$$

y el cuadrado de la distancia escalar entre las medias proyectadas es la distancia de Mahalanobis entre los vectores de medias originales:

$$\left(m_2 - m_1\right)^2 = \left(\mathbf{w}'\left(\mu_2 - \mu_1\right)\right)^2 = \left(\mu_2 - \mu_1\right)'\mathbf{V}^{-1}\left(\mu_2 - \mu_1\right) = D^2$$

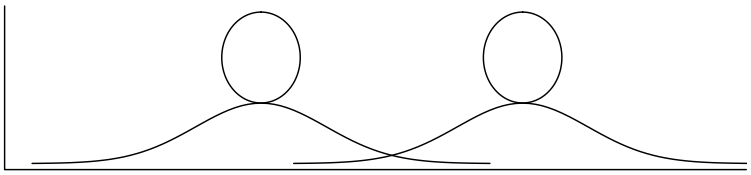


Figura 3.1: Representación de la dirección óptima de proyección para discriminar entre las dos poblaciones.

La variable indicadora z puede interpretarse como una proyección si estandarizamos el vector $\mathbf{w}$. Dividiendo los dos miembros de (3.3) por la norma de $\mathbf{w}$ y llamando $\mathbf{u}$ al vector unitario $\frac{\mathbf{w}}{||\mathbf{w}||}$, la regla de clasificación se convierte en clasificar en P_2 si

$$\mathbf{u}'\mathbf{x} - \mathbf{u}'\mu_1 > \mathbf{u}'\mu_2 - \mathbf{u}'\mathbf{x}$$

donde, al ser $\mathbf{u}$ un vector unitario, $\mathbf{u}'\,\mathbf{x}$ es simplemente la proyección de $\mathbf{x}$ en la dirección de $\mathbf{u}$ y $\mathbf{u}'\mu_1$ y $\mathbf{u}'\mu_2$ las proyecciones de las medias poblacionales en esa dirección. En la Figura 3.1 se observa que el hiperplano perpendicular a $\mathbf{u}$ por el punto medio $\frac{\mathbf{u}'(\mu_1+\mu_2)}{2}$ divide el espacio muestral en dos regiones A_1 y A_2 que constituyen la partición óptima buscada. Si $c(1|2)\pi 2 \neq c(2|1)\pi_1$ la interpretación es la misma, pero el hiperplano frontera se desplaza paralelamente a sí mismo, aumentando o disminuyendo la región A_2. La dirección de proyección, $\mathbf{w} = \mathbf{V}^{-1}(\mu_2 - \mu_1)$ tiene una clara interpretación geométrica. Consideremos en primer lugar el caso en que las variables están incorreladas y estandarizadas de manera que $\mathbf{V} = \mathbf{I}$. Entonces, la dirección óptima de proyección es la definida por $\mu_2 - \mu_1$. En el caso general, la dirección de proyección puede calcularse en dos etapas: primero, se estandarizan las variables de forma multivariante, para pasar a variables incorreladas con varianza unidad; segundo, se proyectan los datos transformados sobre la dirección que une las medias de las variables estandarizadas. En efecto, el cálculo de $\mathbf{w}'\mathbf{x}$ puede escribirse como:

$$\mathbf{w}'\mathbf{x} = \left[(\mu_2 - \mu_1)'\,\mathbf{V}^{-1/2}\right](\mathbf{V}^{-1/2}\mathbf{x})$$

donde $\mathbf{V}^{-1/2}$ existe si $\mathbf{V}$ es definida positiva. Esta expresión indica que esta operación equivale a:

(1) estandarizar las variables $\mathbf{x}$ pasando a otras $\mathbf{y} = \mathbf{V}^{-1/2}\mathbf{x}$ que tienen como matriz de covarianzas la identidad y como vector de medias $\mathbf{V}^{-1/2}\mu$;

(2) proyectar las variables estandarizadas y sobre la dirección $\mu_2(\mathbf{y}) - \mu_1(\mathbf{y}) = (\mu_2 - \mu_1)'\mathbf{V}^{-1/2}$.

Ejercicio 1:

Suponga que se desea discriminar entre dos poblaciones normales con vectores de medias (0,0) y (1,1) , varianzas (2,4) y coeficiente de correlación lineal $r = 0.8$. Construir la función lineal discriminante e interpretarla.

Respuesta :

$\mu_1 = (0,0)$ y $\mu_2 = (1,1)$ $\sigma^2 = (2,4)$ $r = 0.8$

$$\mathbf{V} = \begin{pmatrix} 2 & 2.2627 \\ 2.2627 & 4 \end{pmatrix}$$

$$\mathbf{V}^{-1} = \frac{1}{2.88}\begin{pmatrix} 4 & -2.2627 \\ -2.2627 & 2 \end{pmatrix}$$

La ecuación (3.3) nos dice que clasificamos $\mathbf{x}$ en P_2 si :

$$\mathbf{w}'\mathbf{x} - \mathbf{w}'\mu_1 > \mathbf{w}'\mu_2 - \mathbf{w}'\mathbf{x}$$

donde

$$\mathbf{w} = \mathbf{V}^{-1}(\mu_2 - \mu_1)$$

Entonces

$$\mathbf{w} = \begin{pmatrix} 1.3889 & -0.7857 \\ -0.7857 & 0.6944 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0.6032 \\ -0.0913 \end{pmatrix}$$

Como $\mu_1 = (0, 0)$ nuestra ecuacion queda:

$$2\mathbf{w}'\mathbf{x} > \mathbf{w}'\mu_2 = 0.5119$$

$$\mathbf{w}'\mathbf{x} - 0.25595 > 0$$

$$0.6032x_1 - 0.0913x_2 - 0.25595 > 0$$

Esto se puede interpretar como que la variable con mayor peso en la discriminación es x_1, es la variable que separa más ambas poblaciones. Como, además, x_1 está muy correlada con x_2, conocida x_1 la otra variable no es tan informativa, lo que explica su bajo peso en la función discriminante.

Ejercicio 2:

Las probabilidades a priori en el Ejercicio 1. son 0.7 para la primera población y 0.3 para la segunda. Calcular la función lineal discriminante en este caso.

Respuesta :

Tenemos que la frontera que divide el conjunto de valores posibles para $\mathbf{x}$ está dada por

$$\mathbf{w}'\mathbf{x} = \mathbf{w}'\frac{\mu_1 + \mu_2}{2} - \log\frac{\pi_2}{\pi_1}$$

$$\mathbf{w}'\mathbf{x} = (0.435, 0.163)\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix} - \log(0.0429)$$

Entonces

$$\mathbf{w}'\mathbf{x} = 0.299 + 0.8473 \Rightarrow \mathbf{w}'\mathbf{x} = 1.1464$$

Por lo tanto, clasificaremos a $\mathbf{x}$ en la población P_2 si :

$$\mathbf{w}'\mathbf{x} - 1.1464 > 0$$

$$0.6032x_1 - 0.0913x_2 - 1.1464 > 0$$

3.4 Cálculo de Probabilidades de error

La utilidad de la regla de clasificación depende de los errores esperados. Como la distribución de la variable $z = \mathbf{w}'\mathbf{x}$ es normal, con media $m_i = \mathbf{w}'\mu_i$ y varianza $D^2 = (m_2 - m_1)^2$, podemos calcular las probabilidades de clasificar erróneamente una observación en cada una de las dos poblaciones. En concreto, la probabilidad de una decisión errónea cuando $\mathbf{x} \in P_1$ es:

$$P(2|1) = P\left\{z \geq \frac{m_1 + m_2}{2} | z \text{ es } N(m_1; D)\right\}$$

y llamando $y = \frac{(z-m_1)}{D}$ a una variable aleatoria $N(0,1)$, y Φ a su función de distribución:

$$P(2|1) = P\left\{y \geq \frac{\frac{m_1+m_2}{2} - m_1}{D}\right\} = 1 - \Phi\left(\frac{D}{2}\right)$$

Análogamente, la probabilidad de una decisión errónea cuando $\mathbf{x} \in P_2$ es:

$$P(1|2) = P\left\{z \leq \frac{m_1 + m_2}{2} | z \text{ es } N(m_2; D)\right\} =$$

$$P\left\{y \leq \frac{\frac{m_1+m_2}{2} - m_2}{D}\right\} = \Phi\left(-\frac{D}{2}\right)$$

y ambas probabilidades de error son idénticas, por la simetría de la distribución normal. Podemos concluir que la regal obtenida hace iguales y mínimas (como vimos en la sección 1.3.3) las probabilidades de error y que los errores de clasificación sólo dependen de las distancias de Mahalanobis entre las medias.

En el ejemplo de clasificar la pintura al pintor A o B, el error esperado de clasificación con esta regla depende de la distancia de Mahalanobis entre las medias que es

$$D^2 = (2. - 2.3, 0.8 - 0.7) \begin{bmatrix} 0.25 & 0.025 \\ 0.025 & 0.01 \end{bmatrix}^{-1} \begin{pmatrix} 2. - 2.3 \\ 0.8 - 0.7 \end{pmatrix} = 2.6133$$

y $D = 1.6166$. La probabilidad de equivocarnos es

$$P(A/B) = 1 - \Phi\left(\frac{1.6166}{2}\right) = 1 - \Phi(0.808) = 1 - 0.7905 = 0.2095$$

De manera que la clasificación mediante estas variables no es muy precisa, ya que podemos tener un 18.94% de probabilidad de error. Calculemos la probabilidad a posteriori de que el cuadro pertenezca al pintor A suponiendo que, a priori, ambos pintores son igualmente probables.

$$P(A/\mathbf{x}) = \frac{1}{1 + \exp(-0.5(0.8133 - 0.52)} = \frac{1}{1.86} = 0.5376$$

Esta probabilidad indica que al clasificar la obra como perteneciente al pintor A existe mucha incertidumbre en la decisión, ya que las probabilidades de que pertenezca a cada pintor son semejantes (0.5376 y 0.4624).

Ejercicio 3:

Discutir cómo varían las probabilidades de error en el ejercicio 1 como función del coeficiente de correlación. ¿Ayuda la correlación a la discriminación?

Respuesta :

En la sección 3.4 tenemos que las probabilidades de clasificar erróneamente una observación en cada una de las dos poblaciones es:

$$P(2|1) = 1 - \Phi(\frac{D}{2})$$

Dónde $D = (m_2 - m_1)$, $m_i = \mathbf{w}'\mu_i$ y $\mathbf{w} = \mathbf{V}^{-1}(\mu_2 - \mu_1)$

Tenemos que:

$$V = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}$$

Como el coeficiente de correlación es $r = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$ la matriz V queda:

$$V_r = \begin{pmatrix} \sigma_X^2 & r\sigma_X\sigma_Y \\ r\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

Y por lo tanto

$$V_r^{-1} = \frac{1}{\sigma_X^2\sigma_Y^2(1-r^2)} \begin{pmatrix} \sigma_Y^2 & -r\sigma_X\sigma_Y \\ -r\sigma_X\sigma_Y & \sigma_X^2 \end{pmatrix}$$

Como en el ejercicio 1, tenemos $\mu_1 = (0,0)$ y $\mu_2 = (1,1)$ $\sigma^2 = (2,4)$, reemplazando

$$V_r^{-1} = \frac{1}{8(1-r^2)} \begin{pmatrix} 4 & -r2.8284 \\ -r2.8284 & 2 \end{pmatrix}$$

$$\mathbf{w}_r = V^{-1}(\mu_2 - \mu_1) = \frac{1}{8(1-r^2)} \begin{pmatrix} 4 - 2.8284r \\ 2 - 2.8284r \end{pmatrix}$$

$$D_r = \mathbf{w}'(\mu_2 - \mu_1) = \frac{1}{8(1-r^2)} 6 - 5.6568r$$

Como el coeficiente de correlación, r, toma valores entre -1 y 1, cuando r tiene a 1 en módulo, entonces D_r tiende a infinito. Volviendo a nuestra formula del error

$$P(2|1) = 1 - \Phi(\frac{D_r}{2})$$

mientras mas correladas nuestras variables, menos probabilidad tenemos de equivocarnos, es decir cuando $|r| \to 1$ la probabilidad de equivocarnos tiende a 0 ($P(2|1) \to 0$)

3.5 Generalización para varias poblaciones Normales

3.5.1 Planteamiento General

La generalización de estas ideas para G poblaciones es simple: el objetivo es ahora dividir el espacio E_x en G regiones $A_1, \ldots, A_g, \ldots, A_G$ tales que si $\mathbf{x}$ pertenece a A_i el punto se clasifica en la población P_i. Supondremos que los costes de clasificación son constantes y no dependen de la población en que se haya clasificado. Entonces, la región A_g vendrá definida por aquellos puntos con máxima probabilidad de ser generados por P_g, es decir donde el producto de la probabilidad a priori y la verosimilitud sean máximas:

$$A_g = \left\{ \mathbf{x} \in E_x | \pi_g f_g(\mathbf{x}) > \pi_i f_i(\mathbf{x}); \forall i \neq g \right\} \tag{3.4}$$

Si las probabilidades a priori son iguales, $\pi_i = G^{-1}$, $\forall i$, y las distribuciones $f_i(\mathbf{x})$ son normales con la misma matriz de varianzas, la condición (3.4) equivale a calcular la distancia de Mahalanobis del punto observado al centro de cada población y clasificarle en la población que haga esta distancia mínima.

Minimizar las distancias de Mahalanobis $(\mathbf{x} - \mu_g)'\mathbf{V}^{-1}(\mathbf{x} - \mu_g)$ equivale, eliminando el término $\mathbf{x}'\mathbf{V}^{-1}\mathbf{x}$ que aparece en todas las ecuaciones, a minimizar el indicador lineal

$$L_g(\mathbf{x}) = -\mu_g'\mathbf{V}^{-1}\mathbf{x} + \frac{1}{2}\mu_g'\mathbf{V}^{-1}\mu_g \tag{3.5}$$

y llamando

$$\mathbf{w}_g = \mathbf{V}^{-1}\mu_g$$

la regla es

$$\min_g \left(\frac{1}{2}\mathbf{w}_g'\mu_g - \mathbf{w}_g'\mathbf{x}\right)$$

Para interpretar esta regla, observemos que la frontera de separación entre dos poblaciones, (ij), vendrá definida por:

$$A_{ij}(\mathbf{x}) = L_i(\mathbf{x}) - L_j(\mathbf{x}) = 0 \tag{3.6}$$

sustituyendo con (3.5) y reordenando los términos se obtiene:

$$A_{ij}(\mathbf{x}) = 2\left(\mu_i - \mu_j\right)' \mathbf{V}^{-1}\mathbf{x} + \left(\mu_i - \mu_j\right)' \mathbf{V}^{-1}\left(\mu_i + \mu_j\right) = 0$$

y llamando

$$\mathbf{w}_{ij} = \mathbf{V}^{-1}\left(\mu_i - \mu_j\right) = \mathbf{w}_i - \mathbf{w}_j$$

la frontera puede escribirse como:

$$\mathbf{w}_{ij}'\mathbf{x} = \mathbf{w}_{ij}'\frac{1}{2}\left(\mu_i + \mu_j\right)$$

Esta ecuación admite la misma interpretación como proyección que en el caso de dos poblaciones. Se construye una dirección $\mathbf{w}_{ij}$ y se proyectan las medias y el punto $\mathbf{x}$ que tratamos de clasificar sobre esta dirección. La región de indiferencia es cuando el punto proyectado está equidistante de las medias proyectadas. En otro caso, asignaremos el punto a la población de cuya media proyectada esté más próxima.

Vamos a comprobar que si tenemos G poblaciones sólo necesitamos encontrar

$$r = \min(G - 1, p)$$

direciones de proyección. En primer lugar observemos que, aunque podemos construir $\begin{pmatrix} G \\ 2 \end{pmatrix} = G(G-1)/2$ vectores $\mathbf{w}_{ij}$ a partir de las G medias, una vez que tenemos $G-1$ vectores los demás quedan determinados por éstos. Podemos determinar los $G-1$ vectores $\mathbf{w}_{i,i+1}$, para $i= 1, \dots, G-1$, y obtener cualquier otro a partir de estas $G-1$ direcciones.

Por ejemplo:

$$\mathbf{w}_{i,i+2} = \mathbf{V}^{-1}\left(\mu_i - \mu_{i+2}\right) = \mathbf{V}^{-1}\left(\mu_i - \mu_{i+1}\right) - \mathbf{V}^{-1}\left(\mu_{i+1} - \mu_{i+2}\right) = \mathbf{w}_{i,i+1} - \mathbf{w}_{i+1,i+2}$$

En conclusión, si $p > G-1$, el número máximo de vectores $\mathbf{w}$ que podemos tener es $G-1$, ya que los demás se deducen de ellos. Cuando $p \leq G-1$, como estos vectores pertenecen a $\mathbb{R}^p$ el número máximo de vectores linealmente independientes es p.

Es importante resaltar que, como es natural, la regla de decisión obtenida cumple la propiedad transitiva. Por ejemplo, si $G = 3$, y obtenemos que para un punto $(\mathbf{x})$

$$D_1^2(\mathbf{x}) > D_2^2(\mathbf{x})$$
$$D_2^2(\mathbf{x}) > D_3^2(\mathbf{x})$$

entonces forzosamente debemos concluir que $D_1^2(\mathbf{x}) > D_3^2(\mathbf{x})$ y esta será el resultado que obtendremos si calculamos estas distancias, por lo que el análisis es coherente. Además, si $p = 2$, cada una de las tres ecuaciones $A_{ij}(\mathbf{x}) = 0$ será una recta y las tres se cortarán en el mismo punto. En efecto, cualquier recta que pase por el punto de corte de las rectas $A_{12}(\mathbf{x}) = 0$ y $A_{23}(\mathbf{x}) = 0$ tiene la expresión

$$a_1 A_{12}(\mathbf{x}) + a_2 A_{23}(\mathbf{x}) = 0$$

ya que si $\mathbf{x}_0^*$ es el punto de corte como $A_{12}(\mathbf{x}^*) = 0$, por pertenecer a la primera recta, y $A_{23}(\mathbf{x}^*) = 0$, por pertenecer a la segunda, pertenecerá a la combinación lineal. Como, según (3.6), $A_{13}(\mathbf{x}) = L_1(\mathbf{x}) - L_3(\mathbf{x}) = L_1(\mathbf{x}) - L_2(\mathbf{x}) + L_2(\mathbf{x}) - L_3(\mathbf{x})$, tenemos que

$$A_{13}(\mathbf{x}) = A_{12}(\mathbf{x}) + A_{23}(\mathbf{x})$$

y la recta $A_{13}(\mathbf{x})$ debe siempre pasar por el punto de corte de las otras dos.

Ejercicio 4:

Se desea discriminar entre tres poblaciones normales con vectores de medias (0,0), (1,1) y (0,1) con varianzas (2,4) y coeficiente de correlación lineal $r = 0.5$. Calcular y dibujar las funciones discriminantes y hallar su punto de corte.

Respuesta :

$$\mu_1 = (0,0)\ ,\ \mu_2 = (1,1)\ ,\ \mu_3 = (0,1)\ ,\ \sigma^2 = (2,4)\ ,\ r = 0.5$$

Tenemos que la fórmula del coeficiente de correlación es:

$$r = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Entonces,

$$\sigma_{XY} = r\sigma_X\sigma_Y$$

Además, la matriz de covarianzas es

$$\mathbf{V} = \begin{pmatrix} 2 & 1.4142 \\ 1.4142 & 4 \end{pmatrix}$$

$$\mathbf{V}^{-1} = \frac{1}{6}\begin{pmatrix} 4 & -1.4142 \\ -1.4142 & 2 \end{pmatrix}$$

$$L_1(\mathbf{x}) = -2\mu_1'\mathbf{V}^{-1}\mathbf{x} + \mu_1'\mathbf{V}^{-1}\mu_1$$

Reemplazando :

$$L_1(\mathbf{x}) = -2\begin{pmatrix} 0 & 0 \end{pmatrix}\begin{pmatrix} 0.6667 & -0.2357 \\ -0.2357 & 0.3333 \end{pmatrix}\mathbf{x} + \begin{pmatrix} 0 & 0 \end{pmatrix}\begin{pmatrix} 0.6667 & -0.2357 \\ -0.2357 & 0.3333 \end{pmatrix}\begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Luego,

$$L_1(\mathbf{x}) = 0$$

También tenemos

$$L_2(\mathbf{x}) = -2\mu_2'\mathbf{V}^{-1}\mathbf{x} + \mu_2'\mathbf{V}^{-1}\mu_2$$

Reemplazando :

$$L_2(\mathbf{x}) = -2\begin{pmatrix} 1 & 1 \end{pmatrix}\begin{pmatrix} 1.3889 & -0.7857 \\ -0.7857 & 0.6944 \end{pmatrix}\mathbf{x} + \begin{pmatrix} 1 & 1 \end{pmatrix}\begin{pmatrix} 1.3889 & -0.7857 \\ -0.7857 & 0.6944 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Hacemos el cálculo:

$$L_2(\mathbf{x}) = (-0.8619, -0.1953)\mathbf{x} + 0.5286$$
$$L_2(\mathbf{x}) = -0.8619x_1 - 0.1953x_2 + 0.5286$$

Y por último tenemos:

$$L_3(\mathbf{x}) = -2\mu_3'\mathbf{V}^{-1}\mathbf{x} + \mu_3'\mathbf{V}^{-1}\mu_3$$

Reemplazamos :

$$L_3(\mathbf{x}) = -2\begin{pmatrix}0 & 1\end{pmatrix}\begin{pmatrix}1.3889 & -0.7857\\ -0.7857 & 0.6944\end{pmatrix}\mathbf{x} + \begin{pmatrix}0 & 1\end{pmatrix}\begin{pmatrix}1.3889 & -0.7857\\ -0.7857 & 0.6944\end{pmatrix}\begin{pmatrix}0\\1\end{pmatrix}$$

Hacemos el cálculo :

$$L_3(\mathbf{x}) = (0.4714, -0.6667)\mathbf{x} + 0.3333$$

$$L_3(\mathbf{x}) = 0.4714x_1 - 0.6667x_2 + 0.3333$$

La ecuación (3.6) nos dice que la frontera de separacioón entre dos poblaciones (i, j) viene dada por :

$$A_{i,j}\mathbf{x} = L_i\mathbf{x} - L_j\mathbf{x} = 0$$

Luego

$$A_{1,2}\mathrm{x} = L_1\mathrm{x} - L_2\mathrm{x} = -L_2\mathrm{x} = 0 \Longleftrightarrow 0.8619x_1 + 0.1953x_2 - 0.5286 = 0$$

$\Longleftrightarrow$

$$0.8619x_1 + 0.1953x_2 = 0.5286$$

Análogamente :

$$A_{1,3}\mathrm{x} = -L_3\mathrm{x} = 0 \Longleftrightarrow -0.4714x_1 + 0.6667x_2 = -0.3333$$

Por lo tanto, el punto de corte es

$$x_1 = 0.431$$

$$x_2 = 0.8047$$

3.5.2 Procedimiento operativo

Para ilustrar el procedimiento operativo, supongamos cinco poblaciones con $p > 4$, con lo que existirán cuatro reglas de clasificación independientes y las demás se deducen de ellas. Tenemos dos formas de realizar el análisis. La primera es calcular para las G poblaciones las distancias de Mahalanobis (o lo que es equivalente, las proyecciones (3.5)) y clasificar el elemento en la más próxima. La segunda es hacer el análisis comparando las poblaciones dos a dos. Supongamos que hemos obtenido de las comparaciones 2 a 2 los siguientes resultados: ($i > j$ indica que la población i es preferida a la j, es decir, el punto se encuentra más próximo a la media de la población i que a la de j):

$$1 > 2$$

$$2 > 3$$

$$4 > 3$$

$$5 > 4$$

Las poblaciones 2, 3 y 4 quedan descartadas (ya que $1 > 2 > 3$ y $5 > 4$). La duda no resuelta se refiere a las poblaciones 1 y 5. Construyendo (a partir de las reglas anteriores) la regla para discriminar entre estas dos últimas poblaciones, supongamos que

$$5 > 1$$

y clasificaremos en la población 5.

Cuando $p < G-1$ el máximo número de proyecciones linealmente independientes que podemos construir es p, y éste será el máximo número de variables a definir. Por ejemplo, supongamos que $p = 2$ y $G = 5$. Podemos definir una dirección de proyección cualquiera, por ejemplo

$$\mathbf{w}_{12} = \mathbf{V}^{-1}(\mu_1 - \mu_2)$$

y proyectar todas las medias $(\mu_1, \mu_2, \ldots, \mu_5)$ y el punto $\mathbf{x}$ sobre dicha dirección. Entonces, clasificaremos el punto en la población de cuya media proyectada está más próxima. Ahora bien, es posible que sobre esta dirección coincidan las medias proyectadas de varias poblaciones. Si esto ocurre con, por ejemplo, las μ_4 y μ_5, resolveremos el problema proyectando sobre la dirección definida por otra pareja de poblaciones.

Ejemplo 2: Una máquina que admite monedas realiza tres mediciones de cada moneda para determinar su valor: peso (x_1), espesor (x_2) y la densidad de estrías en su canto (x_3). Los instrumentos de medición de estas variables no son muy precisos y se ha comprobado en una amplia experimentación con tres tipos de

monedas usadas, M_1, M_2, M_3 , que las medidas se distribuyen normalmente con medias para cada tipo de moneda dadas por:

$$\begin{array}{rccc} \mu_1 = & 20 & 8 & 8 \\ \mu_2 = & 19.5 & 7.8 & 10 \\ \mu_3 = & 20.5 & 8.3 & 5 \end{array}$$

y matriz de covarianzas

$$\mathbf{V} = \begin{bmatrix} 4 & 0.8 & -5 \\ 0.8 & .25 & -0.9 \\ -5 & -0.9 & 9 \end{bmatrix}$$

Indicar cómo se clasificaría una moneda con medidas $(22, 8.5, 7)$ y analizar la regla de clasificación. Calcular las probabilidades de error. Aparentemente la moneda a clasificar está más próxima a M_3 en las dos primeras coordenadas, pero más próxima a M_1 por x_3, la densidad de estrías. La variable indicador para clasificar entre M_1 y M_3 es

$$z = (\mu_1 - \mu_3)\,\mathbf{V}^{-1}\mathbf{x} = 1.77x_1 - 3.31x_2 + 0.98x_3$$

la media de esta variable para la primera moneda, M_1,es $1.77 \times 20 - 3.31 \times 8 + 0.98 \times 8 = 16.71$ y para la tercera, M_3, $1.77 \times 20.5 - 3.31 \times 8.3 + 0.98 \times 5 = 13.65$. El punto de corte es la media, 15.17. Como para la moneda a clasificar es

$$z = 1.77 \times 22 - 3.31 \times 8.5 + 0.98 \times 7 = 17.61$$

la clasificaremos como M_1. Este análisis es equivalente a calcular las distancias de Mahalanobis a cada población que resultan ser $D_1^2 = 1.84, D_2^2 = 2.01$ y $D_3^2 = 6.69$. Por tanto clasificamos primero en M_1, luego en M_2 y finalmente como M_3. La regla para clasificar entre la primera y la segunda es

$$z = (\mu_1 - \mu_2)\,\mathbf{V}^{-1}\mathbf{x} = -0.93x_1 + 1.74x_2 - 0.56x_3$$

de estas dos reglas deducimos inmediatamente la regla para clasificar entre la segunda y la tercera, ya que

$$(\mu_2 - \mu_3)\,\mathbf{V}^{-1}\mathbf{x} = (\mu_1 - \mu_3)\,\mathbf{V}^{-1}\mathbf{x} - (\mu_1 - \mu_2)\,\mathbf{V}^{-1}\mathbf{x}$$

Analicemos ahora las reglas de clasificación obtenidas. Vamos a expresar la reglas inicial para clasificar entre M_1 y M_3 para las variables estandarizadas, con lo que se evita el problema de las unidades. Llamando $\tilde{x}_i$ a las variables divididas por sus desviaciones típicas $\tilde{x}_1 = x_1/2$; $\tilde{x}_2 = x_2/0.5$, y $\tilde{x}_3 = x_3/3$, la regla en variables estandarizadas es

$$z = 3.54\tilde{x}_1 - 1.65\tilde{x}_2 + 2.94\tilde{x}_3$$

que indica que las variables con más peso para decidir la clasificación son la primera y la tercera, que son la que tienen mayores coeficientes. Observemos que con variables estandarizadas la matriz de covarianzas es la de correlacón

$$R = \begin{bmatrix} 1 & 0.8 & -0.83 \\ 0.8 & 1 & -0.6 \\ -0.83 & -0.6 & 1 \end{bmatrix}$$

El origen de estas correlaciones entre los errores de medida es que si la moneda adquiere suciedad y aumenta ligeramente su peso, también aumenta su espesor y hace más difícil determinar su densidad de estrías. Por eso hay correlaciones positivas entre peso y espesor, al aumentar el peso aumenta el espesor, pero negativas con las estrías. Aunque la moneda que queremos clasificar tiene mucho peso y espesor, lo que indicaría que pertenece a la clase 3, entonces la densidad de estrías debería medirse como baja, ya que hay correlaciones negativas entre ambas medidas, y sin embargo se mide relativamente alta en la moneda. Las tres medidas son coherentes con una moneda sucia del tipo 1, y por eso se clasifica con facilidad en ese grupo. Vamos a calcular la probabilidad a posteriori de que la observación sea de la clase M_1. Suponiendo que las probabilidades a priori son iguales esta probabilidad ser ·

$$P\left(1/x_0\right) = \frac{\exp\left(-D_1^2/2\right)}{\exp\left(-D_1^2/2\right) + \exp\left(-D_2^2/2\right) + \exp\left(-D_3^2/2\right)}$$

y sustituyendo las distancias de Mahalanobis

$$P\left(1/x_0\right) = \frac{\exp(-1.84/2)}{\exp(-1.84/2) + \exp(-2.01/2) + \exp(-6.69/2)} = 0.50$$

y análogamente $P(2/x_0) = 0.46$, y $P(3/x_0) = 0.04$.

Podemos calcular las probabilidades de error de clasificar una moneda de cualquier tipo en otra clase. Por ejemplo, la probabilidad de clasificar una moneda M_3 con esta regla como tipo M_1 es

$$P(z > 15.17/N(13.64, \sqrt{3.07})) = P\left(y > \frac{15.17 - 13.64}{1.75}\right) = P(y > 0.87) = 0.192$$

como vemos esta probabilidad es bastante alta. Si queremos reducirla hay que aumentar la distancia de Mahalanobis entre las medias de los grupos, lo que supone "aumentar" la matriz $\mathbf{V}^{-1}$ o "reducir" la matriz $\mathbf{V}$. Por ejemplo, si

reducimos a la mitad el error en la medida de las estrías introduciendo medidores más precisos, pero se mantiene las correlaciones con las otras medidas, pasamos a la matriz de covarianzas

$$V_2 = \begin{bmatrix} 4 & 0.8 & -2.5 \\ 0.8 & 0.25 & -0.45 \\ -1 & -0.2 & 2.25 \end{bmatrix}$$

la regla de clasificación entre la primera y la tercera es ahora

$$z = \left(\mu_1 - \mu_3\right) \mathbf{V}^{-1}\mathbf{x} = 3.44x_1 - 4.57x_2 + 4.24x_3$$

y la distancia de Mahalanobis entre las poblaciones 1 y 3 (monedas M_1 y M_3) ha pasado de 3.01 a 12.38, lo que implica que la probabilidad de error entre estas dos poblaciones ha disminuido a $1 - \Phi(\sqrt{12.38}/2) = 1 - \Phi(1.76) = 0.04$ y vemos que la probabilidad de error ha disminuido considerablemente. Podemos así calcular la precisión en las medidas que necesitaríamos para conseguir unas probabilidades de error determinadas.

3.6 Poblaciones desconocidas. Caso general

3.6.1 Regla estimada de clasificación

Vamos a estudiar cómo aplicar la teoría anterior cuando en lugar de trabajar con poblaciones disponemos de muestras. Abordaremos directamente el caso de G poblaciones posibles. Como caso particular, la discriminación clásica es para $G = 2$. La matriz general de datos $\mathbf{X}$ de dimensiones $n \times p$ (n individuos y p variables), puede considerarse particionada ahora en G matrices correspondientes a las subpoblaciones. Vamos a llamar x_{ijg} a los elementos de estas submatrices, donde i representa el individuo, j la variable y g el grupo o submatriz. Llamaremos n_g al número de elementos en el grupo g y el número total de observaciones es:

$$n = \sum_{g=1}^{G} n_g$$

Vamos a llamar $\mathbf{x}'_{ig}$ al vector fila $(1 \times p)$ que contiene los p valores de las variables para el individuo i en el grupo g, es decir, $\mathbf{x}'_{ig} = (x_{i1g}, ..., x_{ipg})$. El vector de medias dentro de cada clase o subpoblación será:

$$\overline{\mathbf{x}}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} \mathbf{x}_{ig}$$

y es un vector columna de dimensión p que contiene las p medias para las observaciones de la clase g. La matriz de varianzas y covarianzas para los elementos de la clase g será:

$$\hat{\mathbf{S}}_g = \frac{1}{n_g - 1} \sum_{i=1}^{n_g} (\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)'$$

donde hemos dividido por n_g - 1 para tener estimaciones centradas de las varianzas y covarianzas. Si suponemos que las G subpoblaciones tienen la misma matriz de varianzas y covarianzas, su mejor estimación centrada con todos los datos será una combinación lineal de las estimaciones centradas de cada población con peso proporcional a su precisión. Por tanto:

$$\hat{\mathbf{S}}_w = \sum_{g=1}^{G} \frac{n_g - 1}{n - G} \hat{\mathbf{S}}_g$$

y llamaremos $\mathbf{W}$ a la matriz de sumas de cuadrados dentro de las clases que viene dada por:

$$\mathbf{W} = (n - G)\hat{\mathbf{S}}_w$$

Para obtener las funciones discriminantes utilizaremos $\overline{\mathbf{x}}_g$ como estimación de μ_g, y $\hat{\mathbf{S}}_w$ como estimación de $\mathbf{V}$. En concreto, suponiendo iguales las probabilidades a priori y los costes de clasificación, clasificaremos al elemento en el grupo que conduzca a un valor mínimo de la distancia de Mahalanobis entre el punto $\mathbf{x}$ y la media del grupo. Es decir, llamando $\widehat{\mathbf{w}}_g = \hat{\mathbf{S}}_w^{-1}\overline{\mathbf{x}}_g$ clasificaremos un nuevo elemento x_0 en aquella población g donde

$$\min_g \left(\mathbf{x}_0 - \overline{\mathbf{x}}_g\right)' \hat{\mathbf{S}}_w^{-1} \left(\mathbf{x}_0 - \overline{\mathbf{x}}_g\right) = \min_g \widehat{\mathbf{w}}_g' \left(\overline{\mathbf{x}}_g - \mathbf{x}_0\right)$$

que equivale a construir las variables indicadoras escalares

$$z_{g,g+1} = \widehat{\mathbf{w}}_{g,g+1}' \mathbf{x}_0 \quad g = 1, ..., G$$

donde

$$\widehat{\mathbf{w}}_{g,g+1} = \hat{\mathbf{S}}_w^{-1}(\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_{g+1}) = \widehat{\mathbf{w}}_g - \widehat{\mathbf{w}}_{g+1}$$

y clasificar en g frente a $g + 1$ si

$$\left|z_{g,g+1} - \widehat{m}_g\right| < \left|z_{g,g+1} - \widehat{m}_{g+1}\right|$$

donde $\widehat{m}_g = \widehat{\mathbf{w}}'_{g,g+1}\overline{\mathbf{x}}_g$.

Conviene antes de construir la regla de clasificación realizar un test de que los grupos son realmente distintos, es decir, que no todas las medias μ_g son iguales. Este contraste puede realizarse siguiendo lo expuesto en la sección 10.7 de Peña (2002).

3.6.2 Cálculo de Probabilidades de error

El cálculo de probabilidades de error podría hacerse sustituyendo los parámetros desconocidos por los estimados y aplicando las fórmulas de la sección 3.4, pero este método no es recomendable ya que va a subestimar mucho las probabilidades de error al no tener en cuenta la incertidumbre de estimación de los parámetros. Un mejor procedimiento, que además no depende de la hipótesis de normalidad, es aplicar la función discriminante a las n observaciones y clasificarlas. En el caso de 2 grupos, obtendríamos la tabla:

	Predicciones	
	P_1	P_2
P_1	n_{11}	n_{12}
P_2	n_{21}	n_{22}

donde n_{ij} es el número de datos que viniendo de la población i se clasifica en j. El error aparente de la regla es:

$$\text{Error} = \frac{n_{12} + n_{21}}{n_{11} + n_{22} + n_{12} + n_{21}} = \frac{\text{Total mal clasificados}}{\text{Total}}$$

Este método tiende a subestimar las probabilidades de error ya que los mismos datos se utilizan para estimar los parámetros y para evaluar el procedimiento resultante. Un procedimiento mejor es clasificar cada elemento con una regla que no se ha construido usándolo. Para ello, podemos construir n funciones discriminantes con las n muestras de tamaño $n-1$ que resultan al eliminar uno a uno cada elemento de la población y clasificar después cada dato con la regla construida sin él. Este método se conoce como validación cruzada y conduce a una mejor estimación del error de clasificación. Si el número de observaciones es muy alto, el coste computational de la validación cruzada es alto y una solución más rápida es subdividir la muestra en k grupos iguales y realizar la validación cruzada eliminado en lugar de una observación uno de estos grupos.

En las siguientes secciones vamos a realizar dos ejemplos computacionalmente, los resolveremos con los procedimientos vistos en este capítulo y los compararemos con los resutados arrojados por la función “lda” del paquete MASS (Ripley et al. (2013)).

Cuadro 3.1: Cuadro de las primeras 6 filas de MEDIFIS.

genero	esta	peso	pie	lonb	anches	diamcra	lrt
0	159	49	36	68	42.0	57	40
1	164	62	39	73	44.0	55	44
0	172	65	38	75	48.0	58	44
0	167	52	37	73	41.5	58	44
0	164	51	36	71	44.5	54	40
0	161	67	38	71	44.0	56	42

3.7 Ejemplo 1: MEDIFIS

Vamos a utilizar los datos de MEDIFIS para clasificar personas por su género conocidas las medidas fisicas de la variables.

MEDIFIS Este conjunto de datos contiene 28 observaciones de 8 variables. Las observaciones corresponden a estudiantes españoles y las variables a sus características físicas. Las variables son: género (0 mujer, 1 hombre), estatura (en cm), peso (en Kgr), longitud de pie (en cm.), longitud de brazo (en cm.), anchura de la espalda (en cm.), diámetro del cráneo (en cm.), longitud entre la rodilla y el tobillo (en cm.).

Para este ejemplo utilizaremos R. Mostramos el código utilizado.

```
# Cargamos los datos
Datos <- read.table("DatosMEDIFIS.txt") # MEDIFIS

# Cambiamos el nombre de las columnas
colnames(Datos) <- c("genero","esta", "peso", "pie",
                     "lonb", "anches", "diamcra", "lrt")
```

El Cuadro 3.1 muestra las primeros 6 observaciones del conjunto de de Datos $MEDIFIS$. Podemos ver qué tipo de variables son sus predictoras (discretas, continuas, cualitativas).

Exploración gráfica de los datos

Graficamos algunas de las variables para ver cómo se comportan por género y si las clases se separan.

Figura 3.2 muestra un diagrama de dispersión de las variables Peso y Estatura. Podemos notar que la variable *estatura* separa bien los grupos y que además, como era de esperarse, las variables *estatura* y *peso* están muy correlacionadas.

En las Figuras 3.3 y 3.4 se ve que a nivel individual, la longitud del pie es la variable que más se diferencia entre los géneros (menor solapamiento entre poblaciones).

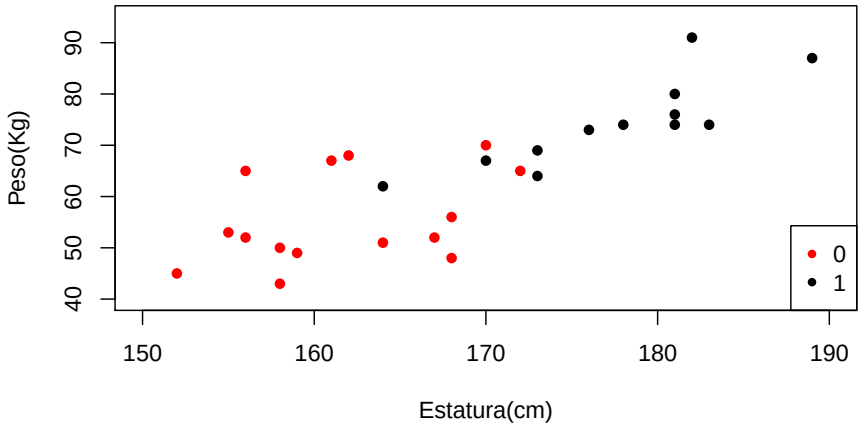


Figura 3.2: Diagrama de dispersión de las variables Estatura y Peso de ambas poblaciones Mujeres (rojo) y Hombres (negro)

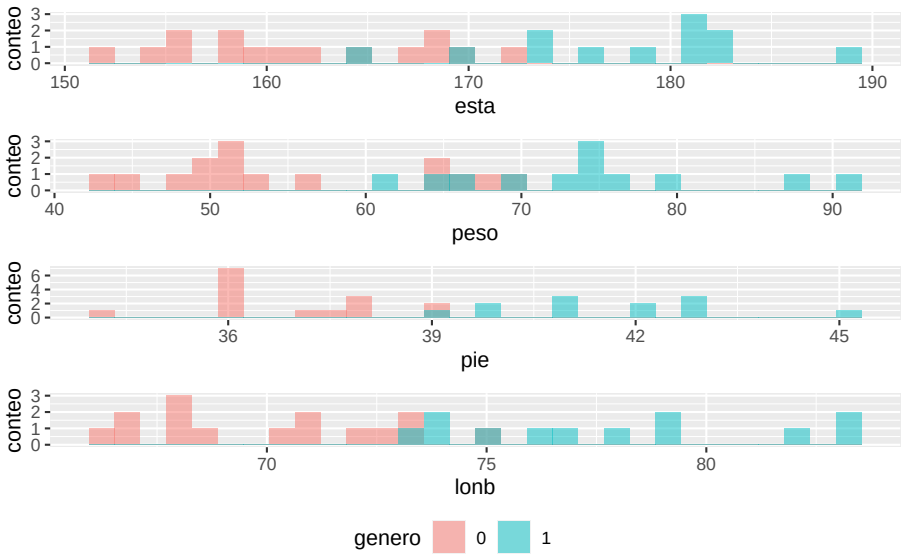


Figura 3.3: Histogramas de las variables estatura, peso, longitud del pie y longitud del brazo separadas cada una por género

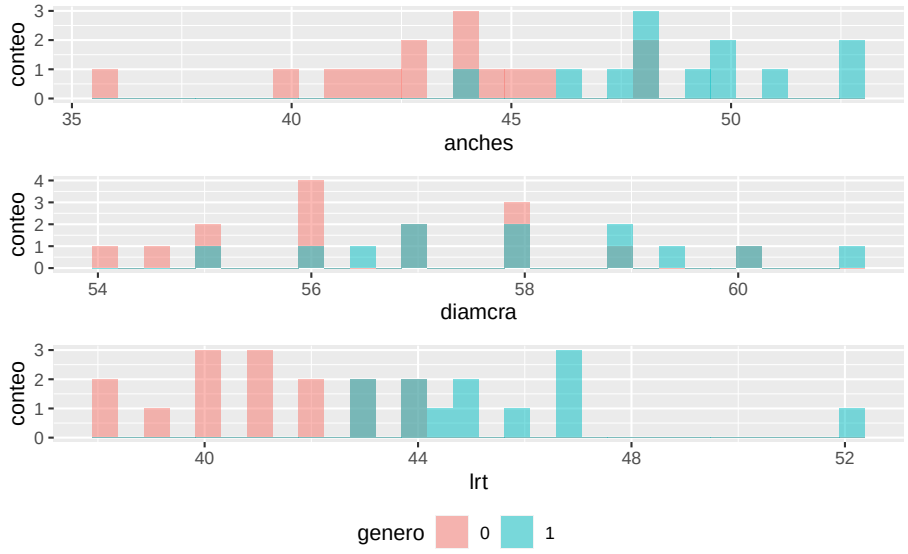


Figura 3.4: Histogramas de las variables ancho de espalda, diámetro del craneo y longitud de rodilla a tobillo separadas cada una por género

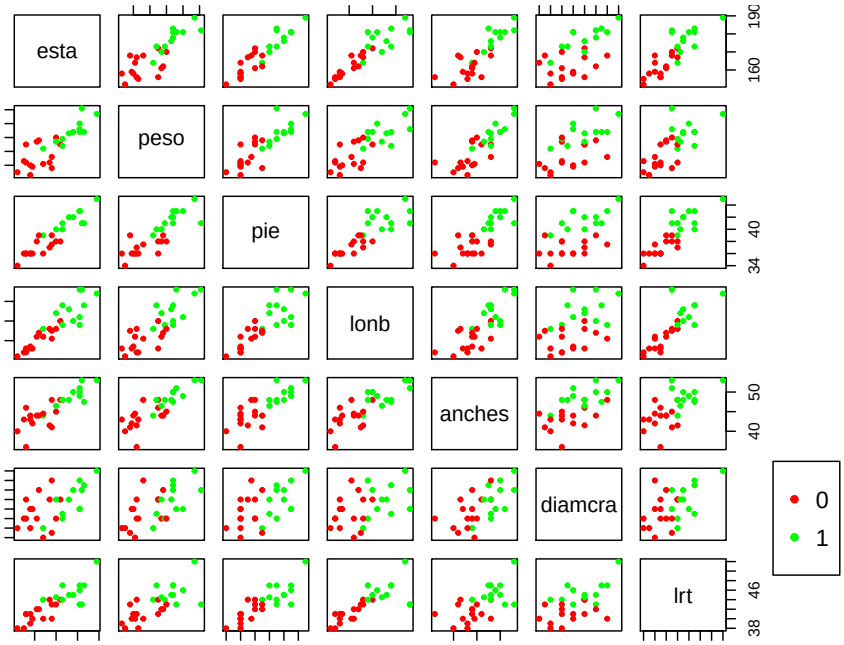


Figura 3.5: Gráficos de dispersión de todas las variables tomadas de a dos

La Figura 3.5 nos muestra que el par de variables pie-lonb y el par pie-lrt parecen separar bien los dos géneros.

3.7.1 Probabilidades a priori

En general, como no se dispone de información sobre la abundancia relativa de las clases a nivel poblacional, se considera como probabilidad previa de cada clase el número de observaciones de la clase entre el número de observaciones totales. Es decir

$$\pi_H = \frac{12}{27} = 0,44$$
$$\pi_M = \frac{15}{27} = 0,56$$

Pero como se considera que la población de hombres es igual a la de las mujeres diremos que

$$\pi_H = \pi_M = 0,5$$

3.7.2 Cálculo de la función discriminante

Mostramos el código para el cálculo de la función discriminante en R:

```
# Identificamos las dos poblaciones.
Hombres <- Datos[Datos$genero == 1, 2:8]
Mujeres <- Datos[Datos$genero == 0,2:8]

# Calculamos las medias de ambas clases.
H_medias <- apply(Hombres, 2, mean)
M_medias <- apply(Mujeres, 2, mean)

# Calculamos las varianzas de ambas clases.
H_varianza <- var(Hombres)
M_varianzas <- var(Mujeres)

# Supondremos que la varianza se mantiene constante en las dos
# clases aplicamos la fórmula vista en la sección 3.6.
n1 <- dim(Mujeres)[1]
n2 <- dim(Hombres)[1]

Sw <- ((n1 - 1) * M_varianzas + (n2 - 1) * H_varianza) / (dim(Datos)[1] - 2)

# Calculamos la inversa de la matriz de covarianzas.
```

```
Sw_inv <- solve(Sw)

# Usamos la fórmula vista en la sección 3.6 para w_{g} con g = 1,2
w_H <- round(Sw_inv %*% H_medias, 4)
w_M <- round(Sw_inv %*% M_medias,4)

# Agergamos el término independiente visto en la sección 3.3
# -1/2 * w´* mu_{g}
f_1 <- rbind(w_H, - t(w_H) %*% H_medias / 2)
f_2 <- rbind(w_M, - t(w_M) %*% M_medias / 2)
```

Las funciones discriminantes $\hat{\mathbf{f}}_g(\mathbf{x}) = \bar{\mathbf{x}}_g' \hat{\mathbf{S}}_w^{-1} \mathbf{x} - \frac{1}{2} \bar{\mathbf{x}}_g' \hat{\mathbf{S}}_w^{-1} \bar{\mathbf{x}}_g$ son:

$\hat{\mathbf{f}}_1 =$

```
##        esta   peso     pie   lonb anches diamcra     lrt
## [1,] -1.303 -4.419 20.0421 9.9804 -2.075  24.357 -4.3652 -1081.715
```

$\hat{\mathbf{f}}_2 =$

```
##         esta    peso     pie  lonb  anches diamcra     lrt
## [1,] -0.9895 -4.3981 17.7293 9.503 -2.5148 25.0768 -4.7245 -1015.588
```

La diferencia entre estas dos funciones proporciona la función lineal discriminante.

```
# Como vimos en la sección 3.6, la fótmula de la función
# discriminante es f_{g,g+1} = f_{g} - f_{g+1}

f_1_2 <- f_1 - f_2
```

$\hat{\mathbf{f}}_{1,2} =$

```
##              [,1]
## esta     -0.31350
## peso     -0.02090
## pie       2.31280
## lonb      0.47740
## anches    0.43980
## diamcra  -0.71980
## lrt       0.35930
##         -66.12769
```

Cuadro 3.2: Matriz de confusión de los datos de Medifis con el método LDA aplicando la función a los mismos datos de entrenamiento

	0	1
0	15	0
1	0	12

Se observa que la variable con mayor peso en la discriminación es la longitud del pie. Como, además, la longitud del pie esta muy correlada con la estatura y la longitud del brazo, conocida la longitud del pie estas variables no son tan informativas, lo que explica su bajo peso en la función discriminante.

Una vez obtenidas las funciones discriminantes, se puede clasificar un nuevo individuo en función de sus medidas. A modo ilustrativo, utilizaremos un dato de nuestra muestra como nueva observación.

```
nueva_observacion <- Datos[1, 2:8]

# Multiplicamos nuestro clasificador por la nueva observación
t(f_1_2) %*% t(cbind(nueva_observacion, 1))
```

```
##              1
## [1,] -9.459293
```

Como el resultado es menor a 0, clasificamos la nueva observación en la población 2, en este caso, en el género Mujeres.

En el Cuadro 3.2 vemos una matriz de confusión aplicando la función discriminante para clasificar los datos muestrales, observamos un porcentaje de éxitos del 100%. Todas las observaciones se clasifican bien.

Acá encontraremos una diferencia cuando utilizamos la función "lda" del paquete MASS, pues los coeficientes que arroja esta función son:

```
library(MASS)
lda(genero ~ . ,data = Datos)$scaling
```

```
##                 LD1
## esta    -0.088142097
## peso    -0.005872569
## pie      0.650276334
## lonb     0.134222958
## anches   0.123645380
## diamcra -0.202387598
## lrt      0.101000944
```

Esta diferencia surge en los distintos enfoques por el cual podemos llegar a los coeficientes de LDA. El que nosotros vimos se basa en las probabilidades a posteriori, el enfoque utilizado por la funcion "lda" del paquete MASS es el de encontrar una dirección α que maximice la distancia entre las medias proyectadas. En la sección 13.5 de Peña (2002) se deduce que hay que encontrar α tal que maximice:

$$\phi = \frac{\left(\alpha' \left(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1\right)\right)^2}{\alpha' S_w \alpha}$$

donde S_w es la matriz de covarianzas ya vista en este trabajo. Lo importante aquí es que la ecuación anterior no depende de la longitud de α ya que entra tanto en el nominador como en el denomidador. Por lo tanto podemos elegir α tal que $\alpha' S_w \alpha = 1$.

```
# En nuestro codigo, después de calcular
b = solve(Sw) %*% (H_medias - M_medias)

# podemos comprobar el valor de b'*S_{w}*b
t(b) %*% Sw %*% b
```

```
##          [,1]
## [1,] 12.64967
```

```
# y es igual a 12.64967. He aquí, su raíz cuadrada es 3.556637,
# y normalizando por ella, llegamos a la salida:lda()
b_lda = b / drop(sqrt(t(b) %*% Sw %*% b))

# Comprobamos que b'*S_{w}*b = 1
drop(t(b_lda) %*% Sw %*% b_lda)
```

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

```
# Y obtenemos la misma función discriminante que lda del paquete MASS
b_lda
```

```
##                [,1]
## esta    -0.088142097
## peso    -0.005872569
## pie      0.650276334
## lonb     0.134222958
## anches   0.123645380
## diamcra -0.202387598
## lrt      0.101000944
```

Cuadro 3.3: Matriz de confusión de los datos de Medifis aplicando LOOCV con el método LDA con la función discriminante calculada a mano

	0	1
0	9	6
1	3	9

Cuadro 3.4: Matriz de confusión de los datos de Medifis aplicando LOOCV con el método LDA del paquete MASS

	1	2
0	13	2
1	2	10

Y estos coeficientes coinciden con los arrojados por la función "lda" del paquete MASS.

En los Cuadros 3.3 y 3.4 vemos una matriz de confusión aplicando validación cruzada (LOOCV) con las funciones dicriminantes hechas a mano y con la función "lda" del paquete MASS. El Cuadro 3.3 supone una proporción de aciertos de 18/27=0.6667. Las observaciones mal clasificadas son las 1, 2, 5, 6, 7, 9, 11, 12, 25. Y en el Cuadro 3.3 supone una proporción de aciertos de 23/27=0.8519. Las observaciones mal clasificadas son las 2, 7, 9, 18.

Vemos que el método de validación cruzada da una idea más realista de la eficacia del procedimiento de clasificación.

3.7.3 Análisis de supuestos de LDA

Vamos a verificar si se cumplen los supuestos que requiere LDA utilizando diferentes métodos gráficos o pruebas de hipótesis. Para tener más detalles de los contrastes que veremos en esta sección ir al Apéndice A.

Análisis de normalidad

Los análisis de normalidad, también llamados contrastes de normalidad, tienen como objetivo analizar cuánto difiere la distribución de los datos observados respecto a lo esperado si procediesen de una distribución normal con la misma media y desviación típica. Pueden diferenciarse dos estrategias: las basadas en representaciones gráficas y en pruebas de hipótesis.

Empezaremos por ver algunos métodos gráficos. Si bien es cierto que los métodos gráficos son más fáciles de interpretar, las pruebas estadísticas nos permiten una mejor generalización de los resultados. El primero que veremos consiste en

representar los datos mediante un histograma y superponer la curva de una distribución normal con la media y desviación estándar iguales a la media y desviación estándar muestrales.

```
# Representación mediante Histograma de cada variable para
#cada tipo de género
par(mfcol = c(4, 4))
for (k in 2:8) {
  j0 <- names(Datos)[k]
  x0 <- seq(min(Datos[, k]), max(Datos[, k]), le = 50)
  for (i in 1:2) {
    i0 <- levels(factor(Datos$genero))[i]
    x <- Datos[Datos$genero == i0, j0]
    hist(x, proba = T, col = grey(0.8), main = paste( i0), xlab = j0, ylab = "Densidad")
    lines(x0, dnorm(x0, mean(x), sd(x)), col = "red", lwd = 2)
  }
}
```

En la Figura 3.6 vemos ese tipo de gráfico para los datos de MEDIFIS por cada variable y por cada género. Podemos ver que la mayoría de los histogramas se aproximan a la forma de la densidad de una normal, también hay variables que a priori cuyas barras de histogramas varían de la curva de densidad normal, como el *pie* en el caso de las *Mujeres* y la longitud entre la rodilla y el tobillo en el caso de los *Hombres*. Si bien mirar estos gráficos no es suficiente para determinar si una variable sigue una distribución normal o no, estos métodos gráficos acompañan muy bien a los contrastes de normalidad que veremos más adelante.

Otro método gráfico que veremos es el gráfico de cuantiles teóricos o también llamado gráfico cuantil-cuantil (q-q plot) y consiste en comparar los cuantiles de la distribución observada con los cuantiles teóricos de una distribución normal con la misma media y desviación estándar que los datos. Cuanto más se aproximen los datos a una normal, más alineados están los puntos entorno a la recta de pendiente uno.

```
# Representación de cuantiles normales de cada variable para cada
#tipo de género
par(mfcol = c(4, 4))
for (k in 2:8) {
  j0 <- names(Datos)[k]
  x0 <- seq(min(Datos[, k]), max(Datos[, k]), le = 50)
  for (i in 1:2) {
    i0 <- levels(factor(Datos$genero))[i]
    x <- Datos[Datos$genero == i0, j0]
    qqnorm(x, main = paste(i0, j0), pch = 19, col = i + 1, xlab = "Cuant. teóricos", ylab = "Cuan
    # los colores 2 y 3 son el rojo y verde
```

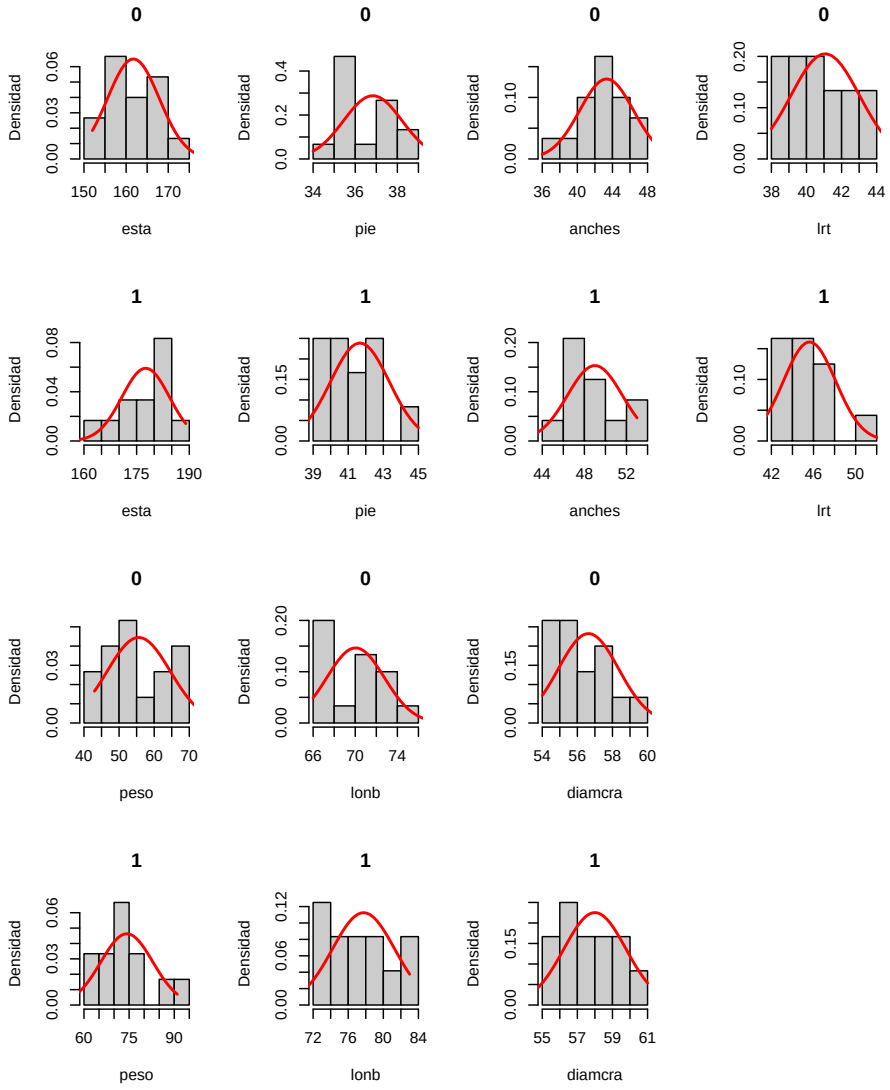


Figura 3.6: Histogramas de variables diferenciadas por género

Cuadro 3.5: Contraste de Normalidad univariante Shapiro-Wilks por cada variable y por cada clase

genero	variable	valor
0	esta	0.5383
1	esta	0.8028
0	peso	0.0845
1	peso	0.5702
0	pie	0.0680
1	pie	0.7868
0	lonb	0.3821
1	lonb	0.2984
0	anches	0.5477
1	anches	0.7361
0	diamcra	0.6951
1	diamcra	0.9909
0	lrt	0.4865
1	lrt	0.0362

```
    qqline(x)
  }
}
```

En la Figura 3.7 vemos los q-q plots por cada variable y por cada género de los datos de MEDIFIS, al igual que en la Figura 3.6 vemos que la mayoría de variables sigue una ditribución normal, ya que los puntos se aproximan a la recta de pendiente uno y que la variable *pie* en las $Mujeres$ parece alejarse una distribución normal.

Para estar más seguros de estas afirmaciones, procederemos a hacer pruebas de hipótesis a cada variable por cada género.

La normalidad de las distribuciones univariantes puede contrastarse con los contrastes χ^2, Kolmogorov-Smirnov, Shapiro y Wilk, o con los contrastes basados en coeficientes de asimetría y curtosis, que pueden consultarse en el capítulo 12 de Peña (2001) (o en el apéndice de este trabajo).

Como tenemos 28 obseraciones, usaremos el test de Shapiro-Wilk que se emplea para contrastar normalidad cuando el tamaño de la muestra es menor de 50. Para muestras grandes es equivalente al test de Kolmogorov-Smirnov.

Siendo la hipótesis nula que la población está distribuida normalmente y dado un nivel de significación estadística α, si el $p - valor$ es menor a α entonces la hipótesis nula es rechazada (se concluye que los datos no vienen de una

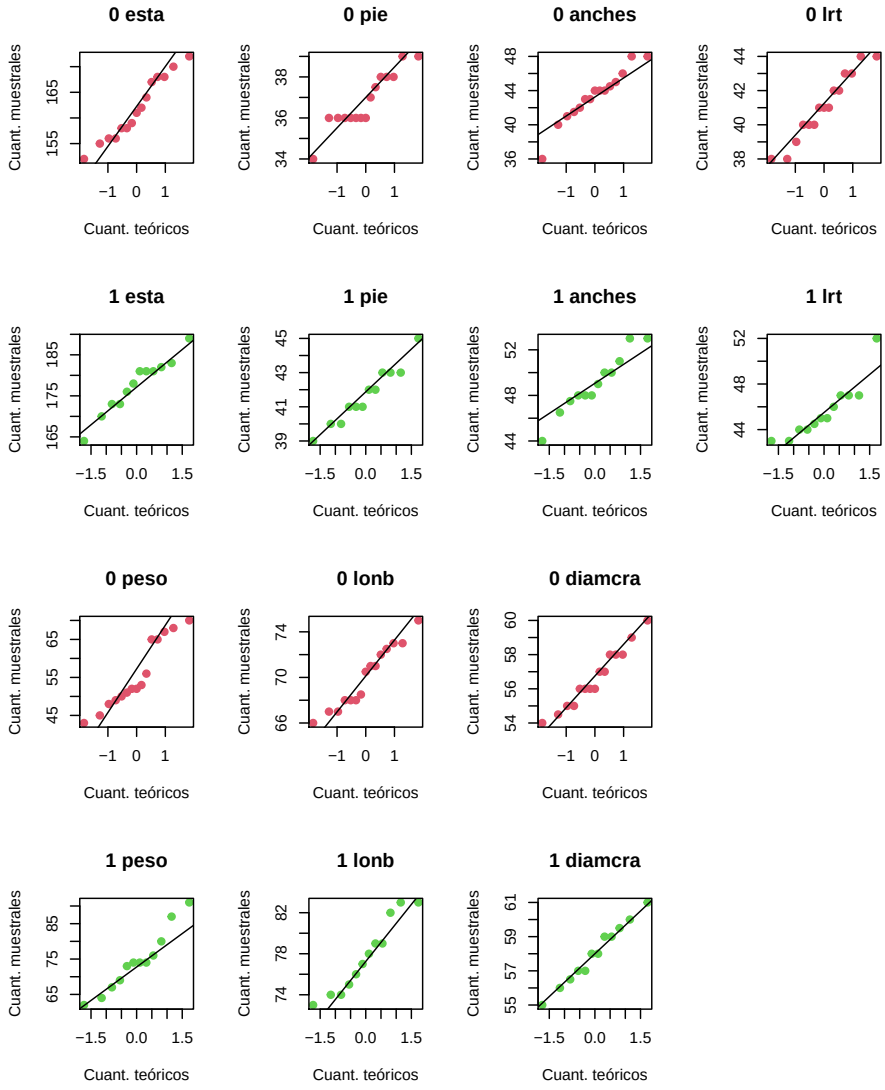


Figura 3.7: Gráficos cuantil cuantil de las variables de MEDIFIS

distribución normal). Si el $p-valor$ es mayor a α, se concluye que no se puede rechazar dicha hipótesis.

En el Cuadro 3.5 vemos los $p-valores$ de las pruebas de normaliad de Shapiro-Wilks de todas las variables de ambas clases. Consideramos $\alpha = 0.05$ el nivel de significación para estos contrastes. A este nivel de significación, observamos que solo hay evidencias suficientes para rechazar la hipótesis de que la variable lrt tiene distribución normal, ya que el $p-valor$ de la prueba resulta 0.0362. Para todas las otras variables, no se rechaza la hipótesis de normalidad.

Luego de haber analizado la normalidad en cada una de las variables por género, vamos a analizar la normalidad multivariante por género, es decir veremos si cada población sigue una distribución normal multivariante. Al observar la normalidad puede suceder que, si los datos tienen una distribución normal multivariante, entonces, cada una de las variables tiene una distribución normal univariante, pero lo opuesto no tiene que ser verdad. Por lo tanto, el control de gráficos y pruebas univariadas podría ser muy útil para diagnosticar el motivo de la desviación.

El paquete Korkmaz et al. (2014) (lo llamaremos MVN) contiene funciones que permiten realizar los tres contrastes de hipótesis comúnmente empleados para evaluar la normalidad multivariante (Mardia, Henze-Zirkler y Royston) y también funciones para identificar valores atípicos que puedan influenciar en el contraste. Ver para más detalles, Apendice A.

Empezaremos analizando los valores atípicos multivariados ya que son la razón mas común para violar la suposición de normalidad multivariada. En otras palabras, la suposición de normalidad multivariada requiere la ausencia de valores atípicos. Por lo tanto, es crucial verificar si los datos tienen valores atípicos antes de comenzar con el análisis multivariado.

En la Figura 3.8 vemos que hay 4 datos atípicos en la población de mujeres ya que su distancia de Mahalanobis ($D_i^2 = (x_i - \mu_M)'\Sigma^{-1}(x_i - \mu_M)$ donde μ_M es la media de la población de mujeres e i es el índice de la observación) es mayor al cuantil $0,975$ de la distribución Chi-cuadrado con 7 grados de libertad ($\chi_7^2(0,975)$).

Al igual que en el caso univariado, existe una prueba gráfica cuantil-cuantil para la normalidad multivarariante, consiste en calcular la distancia de Mahalanobis y compararlos con los cuantiles teóricos de una distribución χ_7^2. Recordemos que la hipótesis nula consiste en $H_0 : D^2 = (\mathbf{x}-\mu_M)'\Sigma^{-1}(\mathbf{x}-\mu_M)$ sigue una distribución χ_7^2 o equivalentemente $H_0 : \mathbf{x}$ sigue una distribución normal multivariada (donde $\mathbf{x}$ es una muestra aleatoria de una población $p-dimensional$).

En la Figura 3.9 vemos que los puntos no se alejan mucho de la recta de pendiente uno, por lo que parecería que ambas poblaciones siguen una distribución normal multivariada.

Para estar mas seguros de esto, procederemos a hacer un contraste de

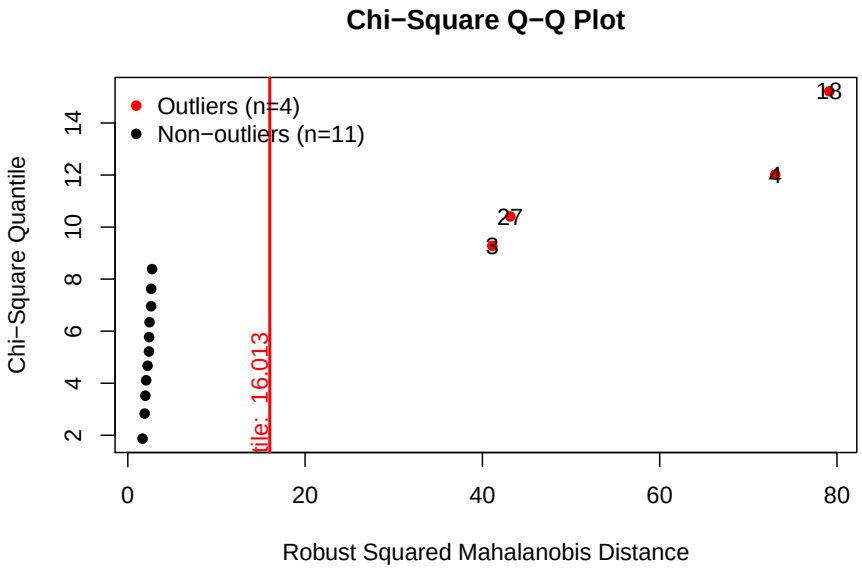


Figura 3.8: Prueba Chi-Cuadrado para valores atípicos en la población de mujeres

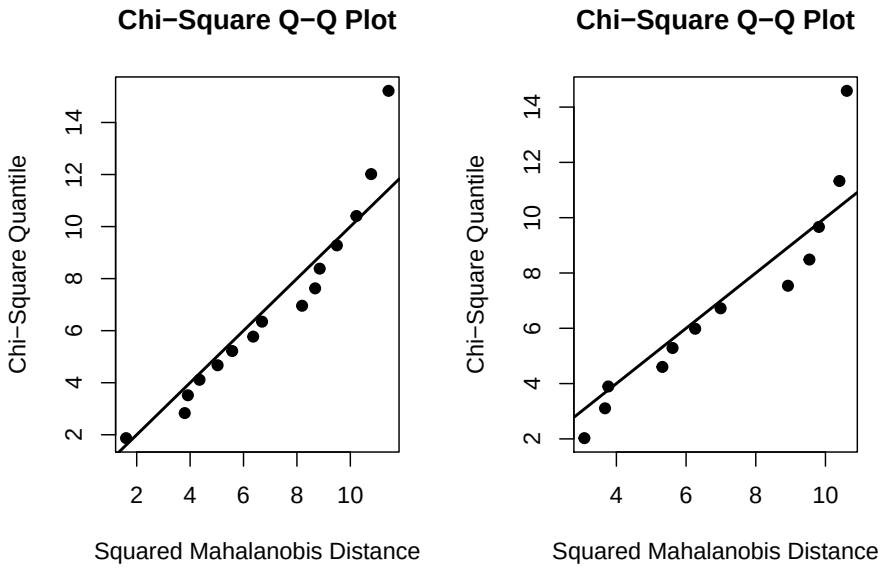


Figura 3.9: Prueba chi-cuadrado normalidad multivariada por género

Cuadro 3.6: Prueba de normalidad multivariada de Royston en mujeres

Test	H	p value	MVN
Royston	7.3753	0.2796	YES

normalidad con los estadísticos de Royston y Henze-Zirkler ya que ambas pruebas se encuentran en el paquete Korkmaz et al. (2014).

Recordemos que para ambos contrastes la hipótesis nula es H_0 : $\mathbf{x}$ sigue una distribución normal multivariada.

El Cuadro 3.6 muestra la prueba Royston para normalidad multivariente para la poblacion de *Mujeres*, en la columna 3 muestra un $p - valor$ de $0.2796 > 0.05$ que es nuestro nivel de significancia, por lo tanto a pesar de los 4 datos atípicos podemos decir que no encontramos evidencia de falta de normalidad multivariante. Además la última columna indica si el conjunto de datos sigue una normalidad multivariante o no (es decir, YES o NO) al nivel de significancia 0.05.

Análogamente, en el Cuadro 3.7 vemos que la población hombres también sigue una distribución normal multivariante para la prueba de Royston.

Cuadro 3.7: Prueba de normalidad multivariada de Royston en hombres

Test	H	p value	MVN
Royston	5.904	0.4581	YES

Cuadro 3.8: Prueba de normalidad multivariada de Henze-Zirkler en mujeres

Test	HZ	p value	MVN
Henze-Zirkler	0.8633	0.4993	YES

En el Cuadro 3.8 vemos el contraste de Henze-Zirkler para la normalidad multivariada y ya sea por la columna 3 del $p - valor > 0.05$ o la columna 4 que dice que SI, aceptamos la hipótesis nula.

En la población de hombres, al contrario que el Cuadro 3.7 que no rechazaba la hipótesis nula, el Cuadro 3.9 nos dice que la rechaza. Esto se puede deber a como vimos antes en el Cuadro 3.5, la variable *lrt* no se distribuye de forma normal. Veremos que pasa si hacemos el contraste quitando la variable *lrt*. En el Cuadro 3.10 muestra los resultados de esta prueba y vemos que no podemos rechazar la hipótesis nula. El LDA tiene cierta robustez frente a la falta de normalidad multivariante, pero es importante tenerlo en cuenta en la conclusión del análisis. Por lo tanto, en el cálculo de la función lineal discriminante, hemos utilizado todas las variables.

Por último, realizaremos un contraste de homogeneidad de varianzas.

Realizamos el contraste Box M (ver Apéndice A) utilizando la funcion "boxM()" del paquete da Silva (2017) obteniendo la siguiente salida:

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  Datos[, 2:8]
## Chi-Sq (approx.) = 39.871, df = 28, p-value = 0.06789
```

Siendo la hipótesis nula $H_0 : \Sigma_M = \Sigma_H$, siendo Σ_M y Σ_H las matrices de covarianzas de $Mujeres$ y $Hombres$ respectivamente, dado que el $p - valor$ es

Cuadro 3.9: Prueba de normalidad multivariada de Henze-Zirkler en hombres

Test	HZ	p value	MVN
Henze-Zirkler	0.9706	0.01	NO

Cuadro 3.10: Prueba de normalidad multivariada de Henze-Zirkler en hombres sin la variable lrt

Test	HZ	p value	MVN
Henze-Zirkler	0.9035	0.0614	YES

$0.06789 > 0,001$ no rechazamos la hipótesis nula, es decir que se puede aceptar que la matriz de covarianza es igual en los dos grupos.

3.8 Ejemplo 2: MUNDODES

Vamos a estudiar la discriminación geográfica entre los países del mundo del banco de datos MUNDODES. Los 91 países incluidos se han clasificado a priori como del este de Europa (9 países, clase 1), América central y del sur (12 países, clase 2), Europa Occidental mas Canadá y EEUU (18 países, clase 3), Asia (25 países, clase 4) y Africa (27 países, clase 5). La variable PNB se ha expresado en logaritmos neperianos.

MUNDODES Este conjunto de datos consta de 91 observaciones y 6 variables. Las observaciones corresponden a 91 países. Las variables son indicadores de desarrollo. Las seis variables son :

Tasa Nat.: Ratio de natalidad por 1000 habitantes

Tasa Mort: Ratio de mortalidad por 1000 habitantes

Mort.Inf: Mortalidad infantil (por debajo de un año)

Esp.Hom: Esperanza de vida en hombres

Esp.Muj.: Esperanza de vida en mujeres

PNB: Producto Nacional Bruto per cápita

Fuente: "UNESCO 1990 Demographic Year Book" y de "The Annual Register 1992".

```
# Cargamos los datos de MUNDODES
Datos <- read.table("DatosMUNDODES.txt")[,-8] #MUNDODES

# Cambiamos el nombre de las columnas
colnames(Datos) <- c("Pais","TasaNat", "TasaMort",
                     "MortInf", "EspViH", "EspViM", "LPNB")

# La variable PNB se ha expresado en logaritmos neperianos.
Datos[,7] <- log(Datos$LPNB)
```

Cuadro 3.11: Primeras 6 observaciones de los datos de MUNDODES

Pais	TasaNat	TasaMort	MortInf	EspViH	EspViM	LPNB
1	24.7	5.7	30.8	69.6	75.5	6.3969
1	12.5	11.9	14.4	68.3	74.7	7.7187
1	13.4	11.7	11.3	71.8	77.7	7.9997
1	11.6	13.4	14.8	65.4	73.8	7.9302
1	14.3	10.2	16.0	67.2	75.7	7.4325
1	13.6	10.7	26.9	66.5	72.4	7.4025

El Cuadro 3.11 muestra las primeros 6 observaciones del conjunto de Datos *MUNDODES*. Podemos ver qué tipo de variables son sus predictoras (discretas, continuas, cualitativas) País, Tasa de Natalidad, Tasa de Mortalidad, Mortalidad Infantil, Espenza de Vida en Hombres, Espenza de Vida en Mujeres, Producto Nacional Bruto per capita. Usaremos estas variables para clasificar un nuevo País en una de las 5 clases mencionadas anteriormente.

Exploración grafica de los datos

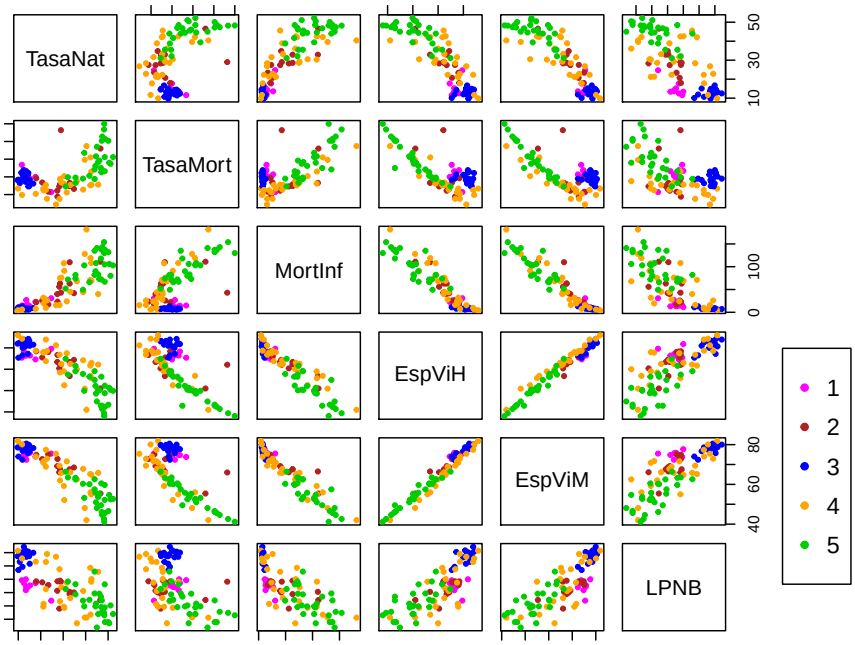


Figura 3.10: Gráficos de dispersión de todas las variables tomadas de a dos

En la Figura 3.10 nos muestra unos diagramas de dispersión de todas las

combinaciones de a 2 de todas las variables. Notemos que, a diferencia del ejemplo de MEDIFIS, como tenemos muchos datos y muchas clases, es muy difícil ver que par de variables separan bien los datos.

3.8.1 Cálculo de la función discriminante

El cálculo de la función discriminante lo realizamos con R:

```
# Codificamos las diferentes clases.
G_1 <- Datos[Datos$Pais == 1, 2:7]
G_2 <- Datos[Datos$Pais == 2, 2:7]
G_3 <- Datos[Datos$Pais == 3, 2:7]
G_4 <- Datos[Datos$Pais == 4, 2:7]
G_5 <- Datos[Datos$Pais == 5, 2:7]

# Calculamos las medias para cada población.
G1_medias <- apply(G_1, 2, mean)
G2_medias <- apply(G_2, 2, mean)
G3_medias <- apply(G_3, 2, mean)
G4_medias <- apply(G_4, 2, mean)
G5_medias <- apply(G_5, 2, mean)

# Calculamos las varianzas para cada población.
S_1 <- var(G_1)
S_2 <- var(G_2)
S_3 <- var(G_3)
S_4 <- var(G_4)
S_5 <- var(G_5)

# Al igual que en el ejemplo anterior, supondremos que la matriz de
#covarianza se mantiene contante en todas las poblaciones.
n1 <- dim(G_1)[1]
n2 <- dim(G_2)[1]
n3 <- dim(G_3)[1]
n4 <- dim(G_4)[1]
n5 <- dim(G_5)[1]

# Calculamos S_w e invertimos.

Sw <- ((n1 - 1)*S_1 + (n2 - 1)*S_2 + (n3 - 1)*S_3 +
         (n4 - 1)*S_4 + (n5 - 1)*S_5) / (dim(Datos)[1] - 5)

Sw_inv <- solve(Sw)

# Calculamos las funciones discriminantes
```

Cuadro 3.12: Funciones discriminantes de cada clase

TasaNat	3.4341	3.7363	3.3751	3.6195	3.9315
TasaMort	9.7586	9.1857	9.6773	8.6879	8.9849
MortInf	1.7345	1.7511	1.7388	1.7107	1.6773
EspViH	-0.1319	0.2815	0.7639	1.7363	0.5993
EspViM	16.9627	16.3425	15.7807	14.3475	15.3422
LPNB	-9.4220	-8.2662	-5.9995	-6.7037	-7.0537
	-689.0495	-681.5261	-688.6183	-640.4625	-645.5401

```
w_1 <- Sw_inv %*% G1_medias
w_2 <- Sw_inv %*% G2_medias
w_3 <- Sw_inv %*% G3_medias
w_4 <- Sw_inv %*% G4_medias
w_5 <- Sw_inv %*% G5_medias

f_1 <- rbind(w_1, - t(w_1) %*% G1_medias / 2)
f_2 <- rbind(w_2, - t(w_2) %*% G2_medias / 2)
f_3 <- rbind(w_3, - t(w_3) %*% G3_medias / 2)
f_4 <- rbind(w_4, - t(w_4) %*% G4_medias / 2)
f_5 <- rbind(w_5, - t(w_5) %*% G5_medias / 2)
```

El Cuadro 3.12 muestra las funciones discriminantes resultantes.

Y para clasificar una nueva observación, evaluaremos la muestra en cada función y la clasificaremos en aquella población donde la función sea máxima.

```
nueva_observacion <- Datos[2, 2:7]

t(f) %*% t(cbind(nueva_observacion, 1))
```

```
##             2
## [1,] 680.3608
## [2,] 675.9143
## [3,] 678.4507
## [4,] 671.4032
## [5,] 667.2314
```

Por lo tanto, clasificamos la nueva observación en la primer población.

Cuadro 3.13: Matriz de confusión aplicando validación cruzada de los datos de MUNDODES con el método LDA calculado a mano

1	2	3	4	5
8	2	1	0	0
1	8	0	2	1
0	1	17	2	0
0	0	0	19	4
0	1	0	2	22

Cuadro 3.14: Matriz de confusión aplicando validación cruzada de los datos de Mundodes con el método LDA del paquete MASS

1	2	3	4	5
7	1	1	0	0
1	6	1	3	1
1	0	17	0	0
0	2	2	18	3
0	1	0	5	21

3.8.2 Evaluación de los errores de clasificación

En el Cuadro 3.13 vemos una matriz de confusión aplicando validación cruzada (LOOCV), que tiene un error de prueba de 0.1868, que es menor al error de prueba utilizando la función lda del paquete MASS, que es 0.2418, y en el Cuadro 3.14 vemos su respectiva matriz de confusión.

3.8.3 Normalidad univariante, normalidad multivariante y homogeneidad de varianza

Al igual que en el ejemplo de MEDIFIS, empezaremos analizando el gráfico de histograma con curva normal. En la Figura 3.11 vemos como la mayoría de las variables se distribuyen de forma normal. En la Figura 3.12 se ven unos resultados muy similares a los de la Figura 3.11, aunque se puede notar que algunas variables se desvían de la recta de pendiente uno. Haremos los contrastes de normalidad de Shapiro-Wilk a cada variable para que no queden dudas. En el Cuadro 3.15 vemos la prueba rechaza la hipótesis nula en la variable TasaNat en los grupos 1 y 5, la variable TasaMort, la prueba rechaza en los grupos 2 y 4, en la variable MortInf la prueba rechaza en los grupos 2, 3 y 4, y por último, en el grupo 3 la rechaza en las variables EspViH y EspViM.

Ahora analizaremos el supuesto de normalidad multivariante y como vimos

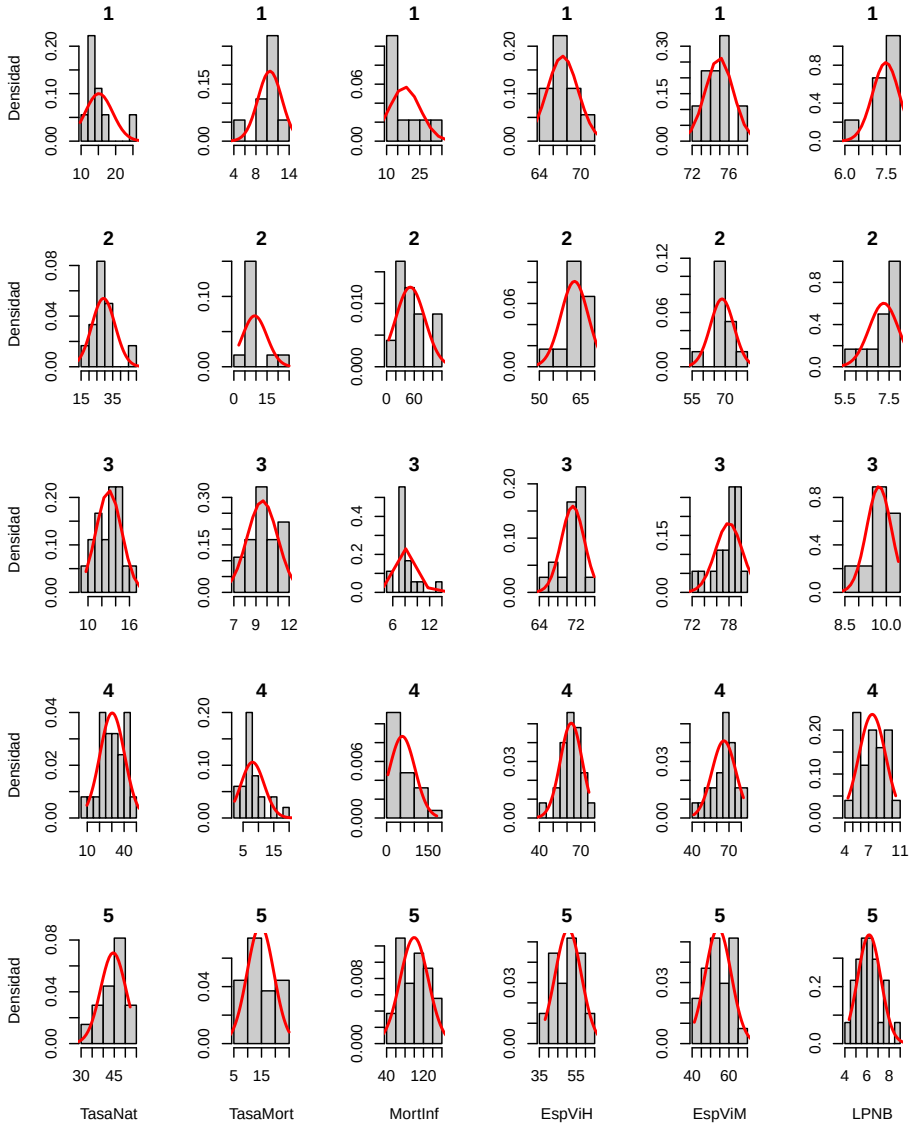


Figura 3.11: Histogramas de variables diferenciadas por grupo con una curva normal superpuesta

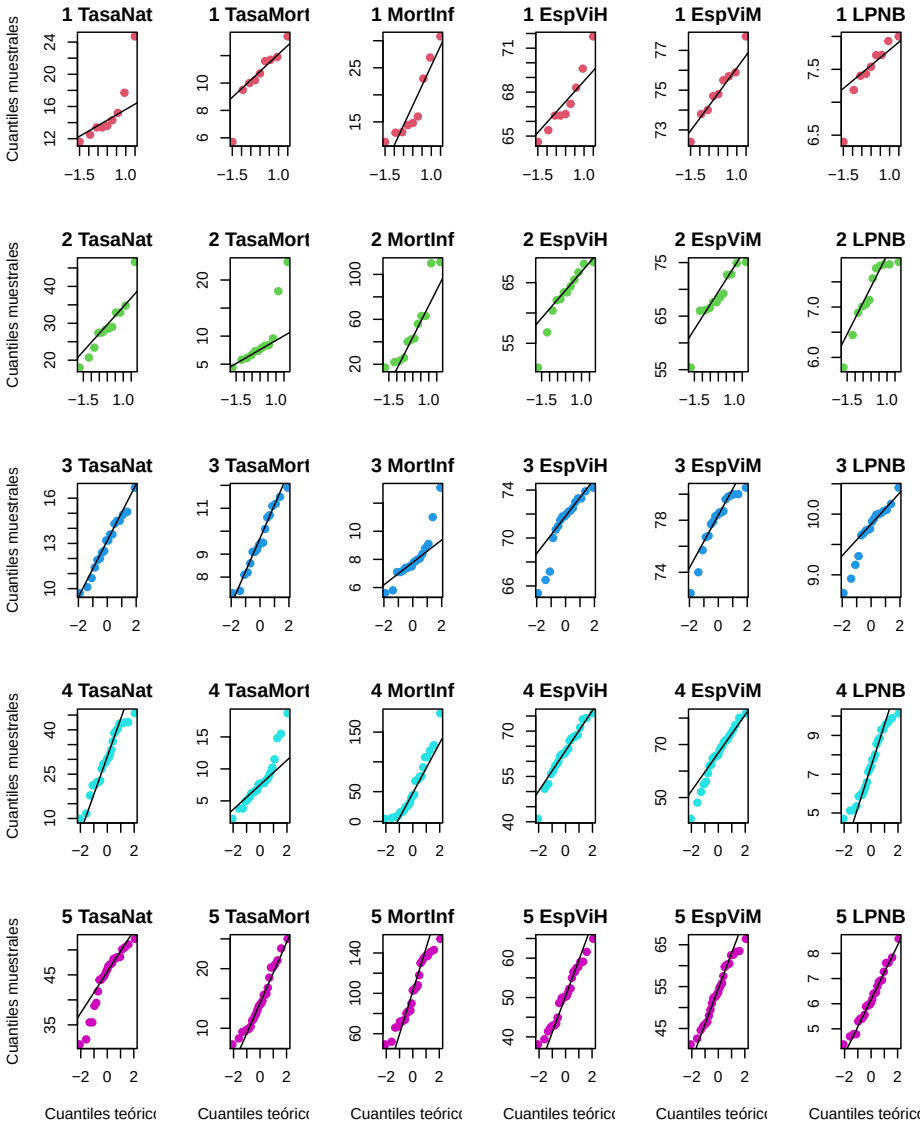


Figura 3.12: Contrastes de normalidad de las variables con el método Q-Q plot

Cuadro 3.15: Contraste de Normalidad univariante Shapiro-Wilks por cada variable y por cada clase

Pais	variable	valor
1	TasaNat	0.0081
2	TasaNat	0.3560
3	TasaNat	0.9675
4	TasaNat	0.2836
5	TasaNat	0.0090
1	TasaMort	0.2129
2	TasaMort	0.0015
3	TasaMort	0.6786
4	TasaMort	0.0125
5	TasaMort	0.2558
1	MortInf	0.0613
2	MortInf	0.0472
3	MortInf	0.0064
4	MortInf	0.0228
5	MortInf	0.1285
1	EspViH	0.4202
2	EspViH	0.1572
3	EspViH	0.0100
4	EspViH	0.3777
5	EspViH	0.6222
1	EspViM	0.9576
2	EspViM	0.0748
3	EspViM	0.0265
4	EspViM	0.5350
5	EspViM	0.5192
1	LPNB	0.1390
2	LPNB	0.0629
3	LPNB	0.0950
4	LPNB	0.1939
5	LPNB	0.8634

anteriormente, empezaremos por la prueba Chi-cuadrado para datos atípicos.

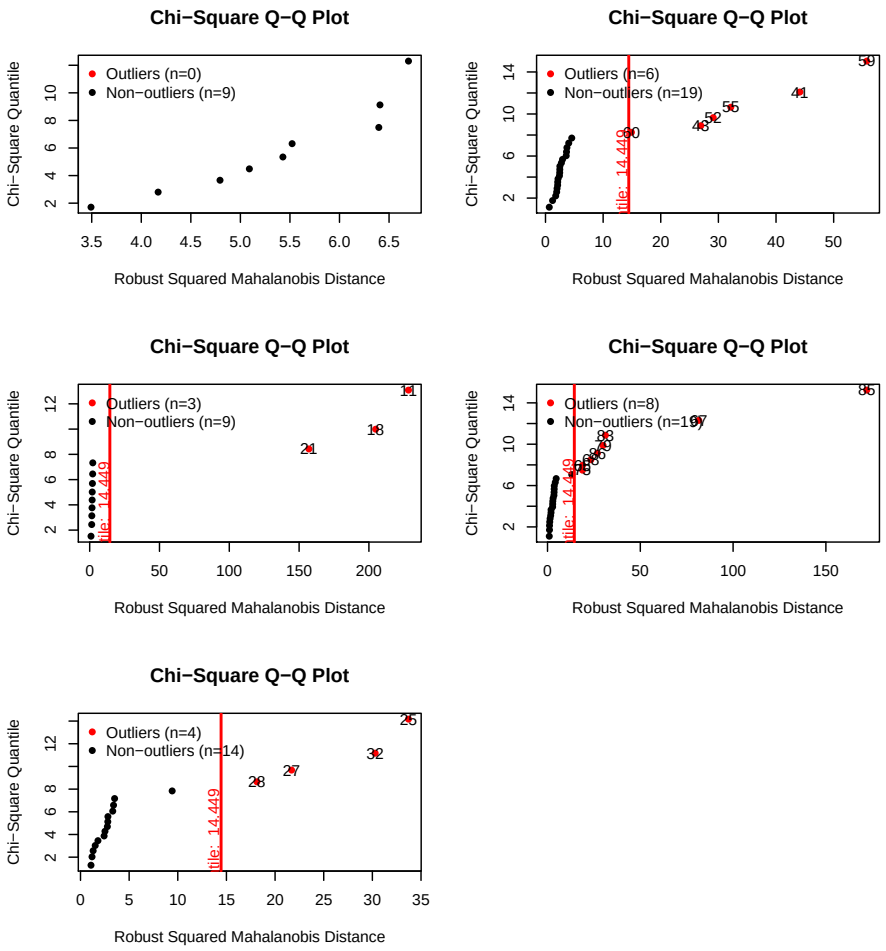


Figura 3.13: Prueba chi-cuadrado para valores atípicos por clase para los datos de MUNDODES

En las Figuras 3.13 y 3.14 podemos ver el comportamiento de las muestras de cada población, como estos gráficos no son suficientes para determinar la normalidad mulitivariante de cada clase, procederemos a los contrastes de normalidad multivariada.

Ambas pruebas muestran evidencias significativas de falta de normalidad multivariante en cada clase, salvo en la población 1 que no rechaza la hipótesis nula en los dos contrastes. Recordemos que el LDA tiene cierta robustez frente a la falta de normalidad multivariante, pero es importante tenerlo en cuenta en

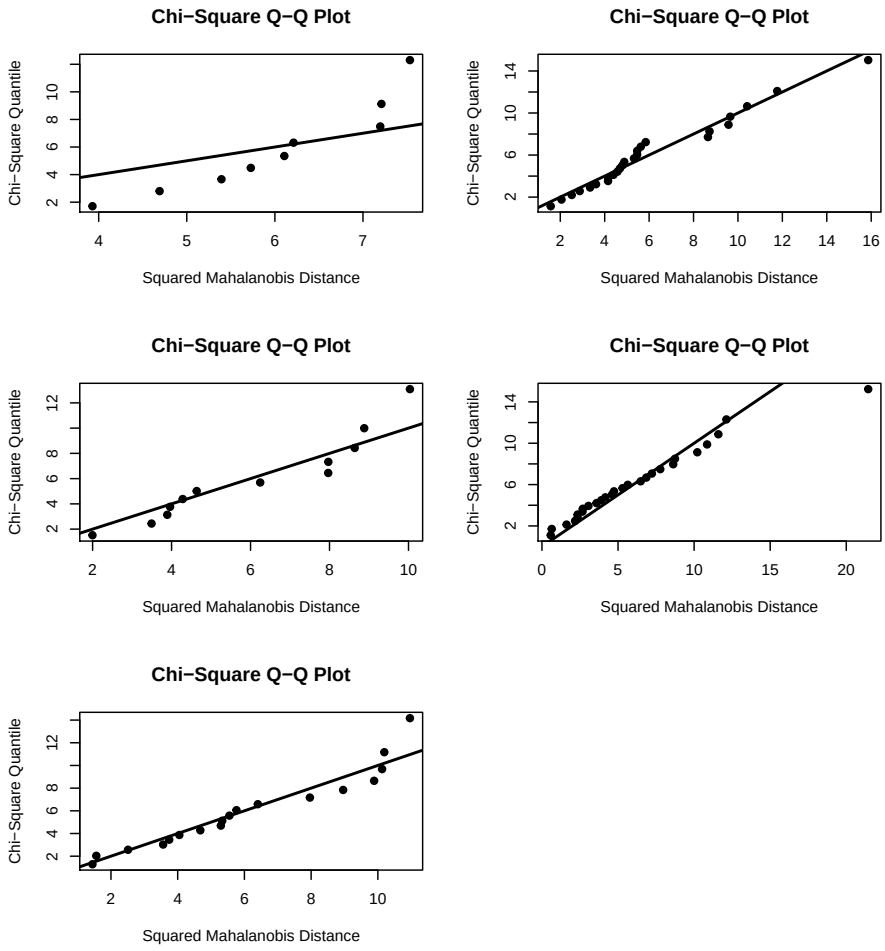


Figura 3.14: Prueba chi-cuadrado normalidad multivariada por clase

Cuadro 3.16: Prueba de normalidad multivariada de Royston en la clase 1

Test	H	p value	MVN
Royston	6.3799	0.0592	YES

Cuadro 3.17: Prueba de normalidad multivariada de Royston en la clase 2

Test	H	p value	MVN
Royston	7.2897	0.0186	NO

Cuadro 3.18: Prueba de normalidad multivariada de Royston en la clase 3

Test	H	p value	MVN
Royston	13.3336	0.0059	NO

Cuadro 3.19: Prueba de normalidad multivariada de Royston en la clase 4

Test	H	p value	MVN
Royston	3.8534	0.0856	YES

Cuadro 3.20: Prueba de normalidad multivariada de Royston en la clase 5

Test	H	p value	MVN
Royston	2.7564	0.1769	YES

Cuadro 3.21: Prueba de normalidad multivariada de Henze-Zirkler en la clase 1

Test	HZ	p value	MVN
Henze-Zirkler	0.8376	0.1987	YES

Cuadro 3.22: Prueba de normalidad multivariada de Henze-Zirkler en la clase 2

Test	HZ	p value	MVN
Henze-Zirkler	0.8457	0.2416	YES

Cuadro 3.23: Prueba de normalidad multivariada de Henze-Zirkler en la clase 3

Test	HZ	p value	MVN
Henze-Zirkler	0.909	0.095	YES

Cuadro 3.24: Prueba de normalidad multivariada de Henze-Zirkler en la clase 4

Test	HZ	p value	MVN
Henze-Zirkler	0.9723	0.0213	NO

Cuadro 3.25: Prueba de normalidad multivariada de Henze-Zirkleren la clase 5

Test	HZ	p value	MVN
Henze-Zirkler	1.062	6e-04	NO

la conclusión del análisis.

Ahora veamos la homogeneidad de varianzas con el test Box M.

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  Datos[, 2:7]
## Chi-Sq (approx.) = 322.67, df = 84, p-value < 2.2e-16
```

El test Box's M muestra evidencias de que la matriz de covarianza no es constante en todos los grupos, lo que a priori descartaría el método LDA. Sin embargo, como el test Box's M es muy sensible a la falta de normalidad multivariante, con frecuencia resulta significativo no porque la matriz de covarianza no sea constante sino por la falta de normalidad, cosa que ocurre para los datos de MUNDODES. Por esta razón se va a asumir que la matriz de covarianza sí es constante y que LDA puede alcanzar una buena precisión en la clasificación. En la evaluación del modelo se verá como de buena es esta aproximación. Además, en las conclusiones se debe explicar la suposición hecha.

Capítulo 4

Análisis Discriminante Cuadrático (QDA)

4.1 Idea intuitiva

Si admitiendo la normalidad de las observaciones la hipótesis de igualdad de varianzas no fuese admisible, el procedimiento de resolver el problema es clasificar la observación en el grupo con máxima probabilidades a posteriori. Esto equivale a clasificar la observación $\mathbf{x}_0$ en la grupo donde se minimice la función :

$$\min_{j\in(1,\ldots,G)} \left[\frac{1}{2}\log\left|\mathbf{V}_j\right| + \frac{1}{2}\left(\mathbf{x}_0-\mu_j\right)' \mathbf{V}_j^{-1}\left(\mathbf{x}_0-\mu_j\right) - \ln\left(C_j\pi_j\right)\right]$$

Cuando $\mathbf{V}_j$ y μ_j son desconocidos se estiman por $\mathbf{S}_j$ y $\overline{\mathbf{x}}_j$ de la forma habitual. Ahora el término $\mathbf{x}_0'\mathbf{V}_j^{-1}\mathbf{x}_0$ no puede anularse, al depender del grupo, y las funciones discriminantes no son lineales y tendrán un término de segundo grado. Suponiendo que los costes de clasificación son iguales en todos los grupos, clasificaremos nuevas observaciones con la regla :

$$\min_{j\in(1,\ldots,G)} \left[\frac{1}{2}\log\left|\widehat{\mathbf{V}}_j\right| + \frac{1}{2}\left(\mathbf{x}_0-\hat{\mu}_j\right)' \widehat{\mathbf{V}}_j^{-1}\left(\mathbf{x}_0-\hat{\mu}_j\right) - \ln\pi_j\right] \tag{4.1}$$

En el caso particular de dos poblaciones y suponiendo las mismas probabilidades a priori clasificaremos una nueva observación en la población 2 si

$$\log\left|\widehat{\mathbf{V}}_1\right| + \left(\mathbf{x}_0-\hat{\mu}_1\right)' \widehat{\mathbf{V}}_1^{-1}\left(\mathbf{x}_0-\hat{\mu}_1\right) > \log\left|\widehat{\mathbf{V}}_2\right| + \left(\mathbf{x}_0-\hat{\mu}_2\right)' \hat{\mathbf{V}}_2^{-1}\left(\mathbf{x}_0-\hat{\mu}_2\right)$$

que equivale a

$$\mathbf{x}_0' \left(\widehat{\mathbf{V}}_1^{-1} - \widehat{\mathbf{V}}_2^{-1} \right) \mathbf{x}_0 - 2\mathbf{x}_0' \left(\widehat{\mathbf{V}}_1^{-1} \hat{\mu}_1 - \widehat{\mathbf{V}}_2^{-1} \hat{\mu}_2 \right) > c \tag{4.2}$$

donde $c = \log \left(\left| \widehat{\mathbf{V}}_2 \right| / \left| \widehat{\mathbf{V}}_1 \right| \right) + \hat{\mu}_2' \widehat{\mathbf{V}}_2^{-1} \hat{\mu}_2 - \hat{\mu}_1' \widehat{\mathbf{V}}_1^{-1} \hat{\mu}_1$. Llamando

$$\widehat{V}_d^{-1} = (\widehat{V}_1^{-1} - \widehat{V}_2^{-1})$$

y

$$\hat{\mu}_d = \widehat{\mathbf{V}}_d \left(\widehat{\mathbf{V}}_1^{-1} \hat{\mu}_1 - \widehat{\mathbf{V}}_2^{-1} \hat{\mu}_2 \right)$$

y definiendo las nuevas variables

$$z_0 = \widehat{V}_d^{-1/2} x_0$$

y llamando $z_0 = (z_{01}, \ldots, z_{0p})'$ y definiendo el vector $m = (m_1, \ldots, m_p)' = \widehat{V}_d^{1/2} (\widehat{V}_1^{-1} \hat{\mu}_1 - \widehat{V}_2^{-1} \hat{\mu}_2)$, la ecuación (4.2) puede escribirse

$$\sum_{i=1}^{p} z_{0i}^2 - 2 \sum_{i=1}^{p} z_{0i} m_i > c$$

Esta es una ecuación de segundo grado en las nuevas variables z_{0i}. Las regiones resultantes con estas funciones de segundo grado son típicamente disjuntas y a veces difíciles de interpretar en varias dimensiones.

4.2 Comparación entre QDA y LDA

El número de parámetros a estimar en el caso cuadrático es mucho mayor que en el caso lineal. En el caso lineal hay que estimar $Gp + p(p+1)/2$ y en el caso cuadrático $G(p + p(p+1)/2)$. Por ejemplo con 10 variables y 4 grupos pasamos de estimar 95 parámetros en el caso lineal a 260 en el caso cuadrático. Este gran número de parámetros hace que, salvo en el caso en que tenemos muestras muy grandes, la discriminación cuadráticas sea bastante inestable y, aunque las matrices de covarianzas sean muy diferentes, se obtengan con frecuencia mejores resultados con la función lineal que con la cuadrática. Un problema adicional con la función discriminante cuadrática es que es muy sensible a desviaciones de la normalidad de los datos. La evidencia disponible indica que la clasificación lineal es en estos casos más robusta. Recomendamos siempre calcular los errores de clasificación con ambas reglas utilizando validación cruzada y en caso de que las diferencias sean muy pequeñas quedarse con la lineal.

En el caso general de poblaciones arbitrarias tenemos dos alternativas:

(a) aplicar la teoría general expuesta en 3.2 y obtener la función discriminante que puede ser complicada.

(b) aplicar la teoría de poblaciones normales, tomar como medida de distancia la distancia de Mahalanobis y clasificar x en la población P_j para la cual la D^2:

$$\mathbf{D}^2 = \left(\mathbf{x} - \overline{\mathbf{x}}_j\right)' \widehat{\mathbf{V}}_j^{-1} \left(\mathbf{x} - \overline{\mathbf{x}}_j\right)$$

es mínima.

4.3 Ejemplo 1

Se dispone de los siguientes datos simulados de dos poblaciones distintas, A y B, con dos variables w y z.

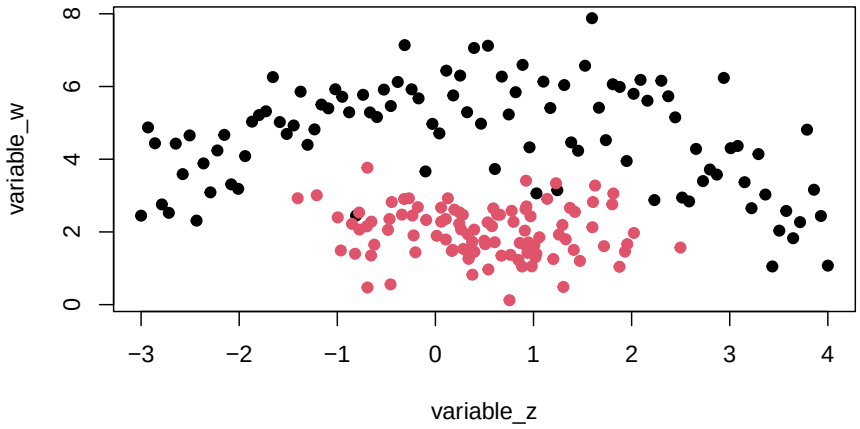


Figura 4.1: Diagrama de dispersión de las variables w y z

En la 4.1 vemos un diagrama de dispersión de las variables ,w y z, los puntos negros pertenecen a la población A y los rojos a la población B. La separación entre los grupos no es de tipo lineal, sino que muestra cierta curvatura. En este tipo de escenarios el método QDA es más adecuado que el LDA.

4.3.1 Exploración gráfica de los datos

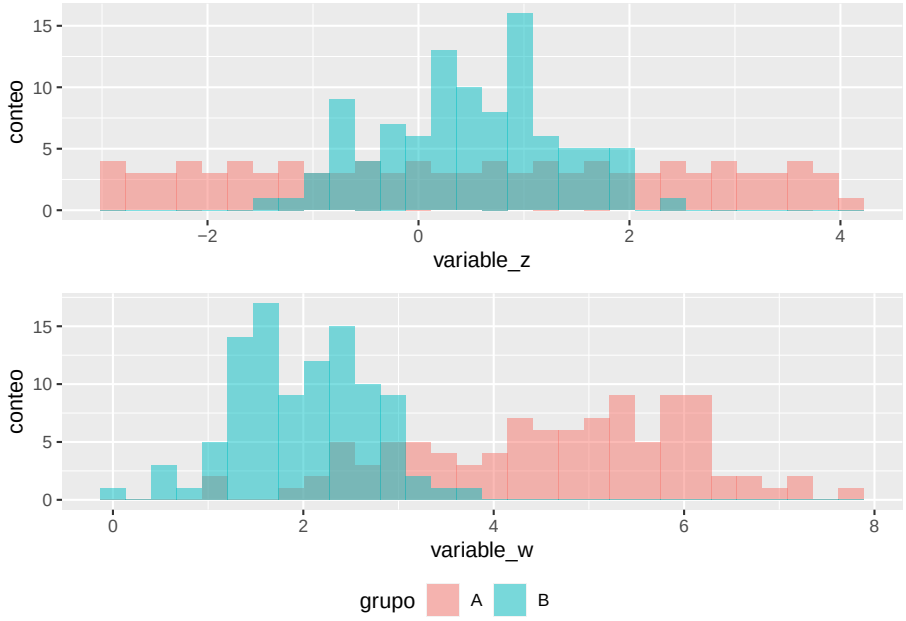


Figura 4.2: Histogramas de las variables w y z separadas cada una por cada clase

En la Figura 4.2 vemos que la variable w permite discriminar entre grupos mejor que la variable z.

4.3.2 Cálculo de la función discriminante.

Desarrollando la ecuación (4.1), deducimos que clasificaremos una nueva observación, $\mathbf{x}_0$, en la población g para la cual

$$f_g(\mathbf{x}_0) = \frac{1}{2}\mathbf{x}_0'\hat{\mathbf{S}}_g^{-1}\mathbf{x}_0 - \mathbf{x}_0'\hat{\mathbf{S}}_g^{-1}\overline{\mathbf{x}}_g + \frac{1}{2}\overline{\mathbf{x}}_g'\hat{\mathbf{S}}_g^{-1}\overline{\mathbf{x}}_g + \frac{1}{2}\log\left|\hat{\mathbf{S}}_g\right| - \log\pi_g$$

sea menor. A diferencia de LDA, es mas difícil mostrar los coeficientes cuadraticos y lineales ya que al estar el termino $\mathbf{x}_0'\hat{\mathbf{S}}_g^{-1}\mathbf{x}_0$ habrá un compromiso entre los elementos fuera de la diagonal de $\hat{\mathbf{S}}_g^{-1}$ y las variables.

En el siguiente código cálculamos las funciones discriminantes y veremos donde clasifica una nueva observación, que para este caso usaremos una muestra de los datos simulados.

Cuadro 4.1: Matriz de confusión aplicando validación cruzada de los datos simulados con el método QDA calculado a mano

1	2
100	0
35	65

```
# Codificamos las diferentes clases.
grupo_A <- datos[datos$grupo == "A", 1:2]
grupo_B <- datos[datos$grupo == "B", 1:2]

# Calculamos las medias para cada población.
m1 <- apply(grupo_A, 2, mean)
m2 <- apply(grupo_B, 2, mean)

# Calculamos las varianzas para cada población.
s1 <- var(grupo_A)
s2 <- var(grupo_B)

# La primera función discriminante
f_1 <- function(x) {

  1/2 * t(x) %*% solve(s1) %*% x - 1/2 * t(x) %*% solve(s1) %*% m1 +
    1/2 * t(m1) %*% solve(s1) %*% m1 + 1/2 * log(det(s1))

}

# La segunda función discriminante
f_2 <- function(x) {

  1/2 * t(x) %*% solve(s2) %*% x - 1/2 * t(x) %*% solve(s2) %*% m2 +
    1/2 * t(m2) %*% solve(s2) %*% m2 + 1/2 * log(det(s2))

}

nueva_observacion <- t(datos[1,1:2])
```

Como $f_1(\mathbf{x}_0) = 6.4564$ y $f_2(\mathbf{x}_0) = 11.8207$, clasificaremos la nueva observación en el grupo A.

Para tener una idea mas realista sobre la eficacia del método de clasificación usamos LOOCV y en el Cuadro 4.1 vemos la matriz de confusión, que tiene un error de prueba de 0.175.

Cuadro 4.2: Matriz de confusión de los datos de Medifis con el método QDA aplicando la función a los mismos datos de entrenamiento

	0	1
0	15	0
1	0	12

Cuadro 4.3: Matriz de confusión de los datos de Medifis aplicando LOOCV con el método QDA

	1	2
0	11	4
1	5	7

4.4 Ejemplo 2: MEDIFIS

En el Cuadro 4.3 vemos que si aplicamos la discriminación cuadrática a los datos de las medidas físicas se obtiene el Cuadro de errores de clasificación por validación cruzada (en el Cuadro 4.2 vemos que sin aplicar validación cruzada se acierta el 100% como en el caso lineal) que supone un porcentaje de aciertos del 67%, menor que en el caso lineal. No hay evidencia de que la discriminación cuadrática suponga ninguna ventaja en este caso.

Capítulo 5

El modelo Logit

5.1 Introducción

La Regresión Logística Simple es un método de regresión que permite estimar la probabilidad de una variable cualitativa binaria en función de una variable cuantitativa. Una de las principales aplicaciones de la regresión logística es la de clasificación binaria, en el que las observaciones se clasifican en un grupo u otro dependiendo del valor que tome la variable empleada como predictor. Por ejemplo, clasificar a un individuo desconocido como hombre o mujer en función del tamaño de la mandíbula.

5.2 Modelos con respuesta cualitativa

Consideremos el problema de la discriminación entre dos poblaciones. Una forma de abordar el problema es definir una variable de clasicación, y, que tome el valor cero cuando el elemento pertenece a la primera población, P_1, y uno cuando pertenece a la segunda, P_2. Entonces, la muestra consistirá en n elementos del tipo $(y_i, \mathbf{x}_i)$, donde y_i es el valor en ese elemento de la variable binaria de clasicación y $\mathbf{x}_i$ un vector de variables explicativas. A continuación, construiremos un modelo para prever el valor de la variable ficticia binaria en un nuevo elemento cuando se conocen las variables $\mathbf{x}$. El primer enfoque simple es formular el modelo de regresión:

$$y = \beta_0 + \beta_1' \mathbf{x} + \mathbf{u} \tag{5.1}$$

y si estimamos los parámetros por mínimos cuadrados este procedimiento es equivalente a la función lineal discriminante de Fisher y es óptimo para clasificar si la distribución conjunta de las variables explicativas es normal

multivariante, con la misma matriz de covarianzas (ver apéndice 13.2 de Peña (2002)). Sin embargo, este modelo presenta problemas de interpretación. Tomando esperanzas en (5.1) para $\mathbf{x} = \mathbf{x}_i$:

$$E[y|\mathbf{x}_i] = \beta_0 + \beta_1'\mathbf{x}_i \tag{5.2}$$

Llamemos p_i a la probabilidad de que y tome el valor 1 (pertenezca a la población P_2) cuando $\mathbf{x} = \mathbf{x}_i$:

$$p_i = P(y = 1|\mathbf{x}_i)$$

la variable y es binomial y toma los valores posibles uno y cero con probabilidades p_i y $1 - p_i$. Su esperanza será :

$$E[y|\mathbf{x}_i] = p_i x1 + (1 - p_i)x0 = p_i \tag{5.3}$$

y de (5.2) y (5.3), concluimos que:

$$p_i = \beta_0 + \beta_1'\mathbf{x}_i \tag{5.4}$$

Esta formulación tiene dos problemas principales:

1. Si estimamos el modelo lineal (5.1), la predicción $\hat{y}_i = \hat{p}_i$ estima, por (5.4), la probabilidad de que un individuo con características definidas por $\mathbf{x} = \mathbf{x}_i$ pertenezca a la segunda población. Sin embargo p_i debe estar entre cero y uno, y no hay ninguna garantía de que la predicción $\hat{y}_i$ verifique esta restricción: podemos obtener probabilidades mayores que la unidad o negativas. Esto no es un problema para clasificar la observación, pero sí lo es para interpretar el resultado de la regla de clasicación.

2. Como los únicos valores posibles de y son cero y uno la perturbación $\mathbf{u}_i$ sólo puede tomar los valores $1 - (\beta_0 + \beta_1 x_i) = 1 - p_i$ y $-\beta_0 - \beta_1' x_i = -p_i$ con probabilidades p_i y $(1 - p_i)$. La esperanza de la perturbación es cero ya que:

$$E[u_i] = p_i(1 - p_i) + (1 - p_i)(-p_i) = 0$$

pero la perturbación no sigue una distribución normal. En consecuencia, los estimadores minimocuadráticos de los coeficientes del modelo (5.1) no serán eficientes. La varianza de u_i es:

$$Var(u_i) = (1 - p_i)^2 p_i + (1 - p_i)p_i^2 = (1 - p_i)p_i$$

y las perturbaciones son heterocedásticas. Para estimar los parámetros del modelo se debera utilizar mínimos cuadrados ponderados.

A pesar de estos dos inconvenientes, este modelo simple estimado por mínimos cuadrados conduce a una buena regla de clasificación, ya que, según la interpretación de Fisher, maximiza la separación entre los grupos, sea cual sea la distribución de los datos. Sin embargo, cuando los datos no son normales, o no tienen la misma matriz de covarianzas, la clasicación mediante una ecuación de relación lineal no es necesariamente óptima.

En el siguiente ejemplo se modela la probabilidad de que una persona sea de un género en función de la longitud del pie, vimos en el capitulo de LDA que puede ser la variable mas significativa (lo veremos tambien en este capítulo mas adelante). En la Figura 5.1 vemos que la relación entre ambas variables estudiadas no es lineal, los puntos rojos representan a las mujeres y los celestes a los hombres.

```
# Cargamos los datos de Medifis vistos anteriormente

datos <- read.table("DatosMEDIFIS.txt")
colnames(datos) <- c("genero","esta", "peso", "pie",
                     "lonb", "anes", "dcra", "lrt")

head(datos)
```

```
##   genero esta peso pie lonb anes dcra lrt
## 1      0  159   49  36   68 42.0   57  40
## 2      1  164   62  39   73 44.0   55  44
## 3      0  172   65  38   75 48.0   58  44
## 4      0  167   52  37   73 41.5   58  44
## 5      0  164   51  36   71 44.5   54  40
## 6      0  161   67  38   71 44.0   56  42
```

```
# Ajuste de un modelo lineal por mínimos cuadrados.
modelo_lineal <- lm(genero ~ pie, data = datos)

# Representación gráfica del modelo.
library(ggplot2)
ggplot(data = datos, aes(x = pie, y = genero)) +
  geom_point(aes(color = as.factor(genero)), shape = 1) +
  geom_smooth(method = "lm", color = "gray20", se = FALSE) +
  theme_bw()  +
  labs(y = "Probabilidad genero") +
  theme(legend.position = "none")
```

Al tratarse de una recta, si por ejemplo, se predice la probabilidad de que una

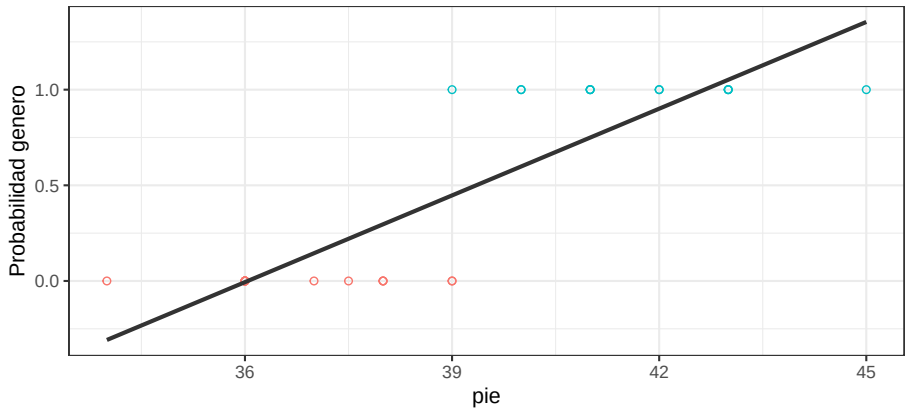


Figura 5.1: Regresión lineal por mínimos cuadrados a los datos de MEDIFIS en el que la variable respuesta es genero y el predictor pie

persona sea de un genero para alguien que tiene una longitud del pie de 45 cm, el valor obtenido es mayor que 1.

```
predict(object = modelo_lineal, newdata = data.frame(pie = 45))
```

```
##        1
## 1.353886
```

Si queremos que el modelo construido para discriminar nos proporcione directamente la probabilidad de pertenecer a cada población, debemos transformar la variable respuesta para garantizar que la respuesta prevista está entre cero y uno. Escribiendo:

$$p_i = F(\beta_0 + \beta_1' \mathbf{x}_i),$$

p_i estará entre cero y uno si escogemos F para que tenga esa propiedad. La clase de funciones no decrecientes acotadas entre cero y uno es la clase de las funciones de distribución, por lo que el problema se resuelve tomando como F cualquier función de distribución. Algunas posibilidades consideradas son:

(1) Tomar como F la función de distribución de una uniforme. Esto equivale a truncar el modelo de regresión, ya que entonces:

$$p_i = 1 \quad \text{si} \quad \beta_0 + \beta_1' \mathbf{x}_i \geq 1$$

$$p_i = \beta_0 + \beta_1' \mathbf{x}_i \quad \text{si} \quad 0 < \beta_0 + \beta_1' \mathbf{x}_i < 1$$
$$p_i = 0 \quad \text{si} \quad \beta_0 + \beta_1' \mathbf{x}_i \leq 0.$$

Esta solución no es sin embargo satisfactoria ni teóricamente (un pequeño incremento de $\mathbf{x}$ produce en los extremos un salto muy grande, cuando sera más lógico una evolución gradual), ni prácticamente: la estimación del modelo es difícil e inestable debido a la discontinuidad.

(2) Tomar como F la función de distribución logística, dada por:

$$p_i = \frac{1}{1 + e^{-\beta_0 - \beta_1' \mathbf{x}_i}}$$

Esta función tiene la ventaja de la continuidad. Además como:

$$1 - p_i = \frac{e^{-\beta_0 - \beta_1' \mathbf{x}_i}}{1 + e^{-\beta_0 - \beta_1' \mathbf{x}_i}} = \frac{1}{1 + e^{\beta_0 + \beta_1' \mathbf{x}_i}}$$

resulta que:

$$g_i = \log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1' \mathbf{x}_i \tag{5.5}$$

que es un modelo lineal en esta transformación que se denomina *logit*. La variable *Logit*, g, representa en una escala logarítmica la diferencia entre las probabilidades de pertenecer a ambas poblaciones, y al ser una función lineal de las variables explicativas nos facilita la estimación y la interpretación del modelo. La Figura 5.2 muestra un gráfico parecido al anterior con la diferencia que el modelo es ajustado por un modelo logístico. Utilizamos la función glm para este fin.

```
# Ajuste de un modelo logístico.
modelo_logistico <- glm(genero ~ pie, data = datos, family = "binomial")

# Representación gráfica del modelo.
ggplot(data = datos, aes(x = pie, y = genero)) +
  geom_point(aes(color = as.factor(genero)), shape = 1) +
  stat_function(fun = function(x){predict(modelo_logistico,
                                          newdata = data.frame(pie = x),
                                          type = "response")}) +
  theme_bw() +
  labs(y = "Probabilidad genero") +
  theme(legend.position = "none")
```

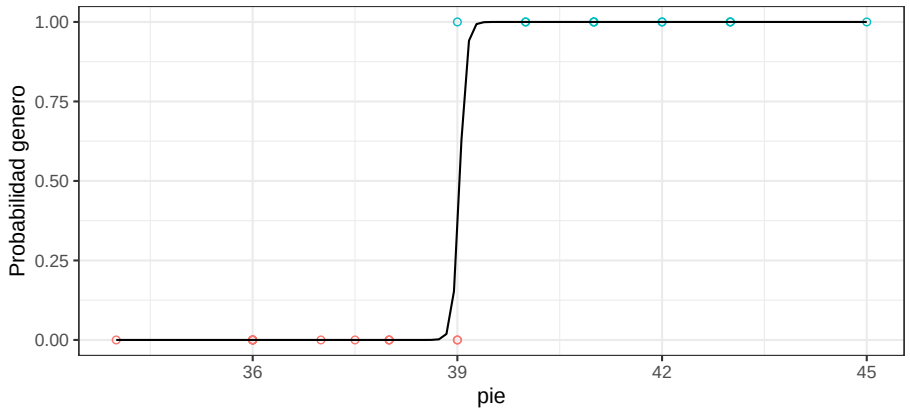


Figura 5.2: Regresión logística a los datos de MEDIFIS en el que la variable respuesta es genero y el predictor pie

(3) Tomar otra distribución, como por ejemplo escoger F igual a la distribución normal estándar. Se obtiene entonces el modelo *probit*, que es muy similar al *logit*, sin tener las ventajas de interpretación del modelo logístico, como veremos a continuación.

5.3 El modelo logit con datos normales

El modelo logit se aplica a una amplia gama de situaciones donde las variables explicativas no tienen una distribución conjunta normal multivariante. Por ejemplo, si algunas son categóricas, podemos introducirlas en el modelo logit mediante variables ficticias como se hace en el modelo de regresión estándar. Una ventaja adicional de este modelo es que si las variables son normales verifican el modelo *logit*. En efecto, supongamos que las variables $\mathbf{x}$ provienen de una de dos poblaciones normales multivariantes con distinta media pero la misma matriz de varianzas covarianzas. Hemos visto en el capítulo anterior (sección 1.3.1) que, suponiendo las probabilidades a priori de ambas poblaciones iguales:

$$p_i = P(y = 1|x_i) = \frac{f_1(x_i)}{f_1(x_i) + f_2(x_i)}$$

y, utilizando la transformación *logit*, (5.5):

$$g_i = \log \frac{f_1(x_i)}{f_2(x_i)} = -\frac{1}{2}(x_i - \mu_1)'V^{-1}(x_i - \mu_1) + \frac{1}{2}(x_i - \mu_2)'V^{-1}(x_i - \mu_2)$$

y simplificando

$$g_i = \frac{1}{2}\left(\mu_2 \mathbf{V}^{-1}\mu_2 - \mu_1 \mathbf{V}^{-1}\mu_1\right) + \left(\mu_1 - \mu_2\right)' \mathbf{V}^{-1}\mathbf{x}_i$$

Por tanto, g_i es una función lineal de las variables $\mathbf{x}$, que es la característica que define el modelo *logit*. Comparando con (5.5) la ordenada en el origen, β_0, es igual

$$\beta_0 = \frac{1}{2}\left(\mu_2 \mathbf{V}^{-1}\mu_2 - \mu_1 \mathbf{V}^{-1}\mu_1\right) = -\frac{1}{2}\mathbf{w}'\left(\mu_1 + \mu_2\right)$$

donde $\mathbf{w} = \mathbf{V}^{-1}(\mu_1 - \mu_2)$, y el vector de pendientes

$$\beta_1 = \mathbf{w}$$

Observemos que la estimación de $\hat{\mathbf{w}}$ mediante el modelo logístico no es eficiente en el caso normal. En efecto, en lugar de estimar los p $(p+1)/2$ términos de la matriz $\hat{\mathbf{V}}$ y los $2p$ de las medias $\overline{\mathbf{x}}_1$ y $\overline{\mathbf{x}}_2$, con el modelo logístico estimamos únicamente $p+1$ parámetros $\beta_0, \beta_1, ..., \beta_p$. En el caso de normalidad se obtiene un mejor procedimiento con la regla de Fisher, que estima $\hat{\mathbf{V}}$, $\overline{\mathbf{x}}_1$ y $\overline{\mathbf{x}}_2$, la distribución completa de las $\mathbf{x}$, mientras que el modelo logístico estima sólo los $p+1$ parámetros de la distribución de y condicionada a $\mathbf{x}$. Como:

$$f(\mathbf{x}, y) = f(y|\mathbf{x})f(\mathbf{x})$$

perdemos información al considerar sólo la condicionada $f(y|\mathbf{x})$ – como hace el modelo logístico – en lugar de la conjunta $f(\mathbf{x}, y)$, que se utiliza en el enfoque del capítulo anterior. Efron (1975) demostró que cuando los datos son normales multivariantes y estimamos los parámetros en la muestra, la función de discriminación lineal de Fisher funciona mejor que regresión logística.

En resumen, en el caso de normalidad la regla discriminante es mejor que el modelo logístico. Sin embargo, la función logística puede ser más eficaz cuando los poblaciones tengan distinta matriz de covarianzas o sean marcadamente no normales. En el campo de la concesión automática de créditos (Credit Socoring) existen numerosos estudios comparando ambos métodos. La conclusión general es que ninguno de los dos métodos supera al otro de manera uniforme y que depende de la base de datos utilizada. Rosenberg and Gleit (1994) y Hand and Henley (1997) han presentado estudios sobre este problema.

5.4 Interpretación del Modelo Logístico

Los parámetros del modelo son β_0, la ordenada en el origen, y $\beta_1 = (\beta_1, ..., \beta_p)$, las pendientes. A veces se utilizan también como parámetros $\exp(\beta_0)$ y $\exp(\beta_i)$, que se denominan los odds ratios o ratios de probabilidades, e indican cuanto se modifican las probabilidades por unidad de cambio en las variables $\mathbf{x}$. En efecto, de (5.5) deducimos que

$$O_i = \frac{p_i}{1-p_i} = \exp(\beta_0) \cdot \prod_{j=1}^{p} \exp(\beta_j)^{x_j}$$

Supongamos dos elementos, i, k, con todos los valores de las variables iguales excepto la variable h y $x_{ih} = x_{jh} + 1$. El cociente de los ratios de probabilidades (odds ratio) para estas dos observaciones es:

$$\frac{O_i}{O_k} = e^{\beta_h}$$

e indica cuanto se modifica el ratio de probabilidades cuando la variable x_h aumenta una unidad. Sustituyendo $\widehat{p}_i = 0.5$ en el modelo *logit*, entonces,

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} = 0$$

es decir,

$$x_{i1} = -\frac{\beta_0}{\beta_1} - \sum_{j=2}^{p} \frac{\beta_j x_{ij}}{\beta_1}$$

y x_{i1} representa el valor de x_1 que hace igualmente probable que un elemento, cuyas restantes variables son $x_{i2}, ..., x_{ip}$, pertenezca a la primera o la segunda población.

5.5 La estimación del modelos logit

5.5.1 Estimación MV

Supondremos una muestra aleatoria de datos $(\mathbf{x}_i, y_i)$, $i = 1, ..., n$. La función de probabilidades para una respuesta y_i cualquiera es:

$$P(y_i) = p_i^{y_i} (1-p_i)^{1-y_i} \quad y_i = 0, 1$$

y para la muestra :

$$P\left(y_1, \ldots y_n\right) = \prod_{i=1}^{n} p_i^{y_i} \left(1 - p_i\right)^{1-y_i}$$

Tomando logaritmos:

$$\log P(\mathbf{y}) = \sum_{i=1}^{n} y_i \log\left(\frac{p_i}{1-p_i}\right) + \sum \log\left(1 - p_i\right) \tag{5.6}$$

la función soporte (de verosimilitud en logaritmos), $L(\theta)$, puede escribirse como

$$\log P(\beta) = \sum_{i=1}^{n} \left(y_i \log p_i + \left(1 - y_i\right) \log\left(1 - p_i\right)\right)$$

donde $\beta' = (\beta_0, \beta_1, \ldots, \beta_p)$ es un vector de $p+1$ componentes, incluyendo la constante β_0 que determina las probabilidades p_i. Maximizar la verosimilitud puede expresarse como minimizar una función que mide la desviación entre los datos y el modelo. En el capítulo 10 de Peña (2002) se define la desviación de un modelo mediante $D(\theta)$ = - $2L(\theta)$ y por tanto la desviación del modelo será:

$$D(\beta) = -2 \sum_{i=1}^{n} \left(y_i \log p_i + \left(1 - y_i\right) \log\left(1 - p_i\right)\right) \tag{5.7}$$

y hablaremos indistintamente de de maximizar el soporte o minimizar la desviación del modelo. Se define la desviación de cada dato (deviance) por:

$$d_i = -2 \left(y_i \log p_i + \left(1 - y_i\right) \log\left(1 - p_i\right)\right) \tag{5.8}$$

y miden el ajuste del modelo al dato $(y_i, \mathbf{x}_i)$. En efecto, observemos en primer lugar que como los $\widehat{p}_i$ son menores que uno, sus logaritmos son negativos, por lo que la desviación es siempre positiva. Además, en el cálculo de la desviación sólo interviene uno de sus dos términos, ya que y_i solo puede valer cero o uno. Entonces:

• Si $y_i = 1$, y la observación pertenece a la segunda población, el segundo término de la desviación es nulo y d_i = -2 $\log p_i$. La observación tendrá una desviación grande si la probabilidad estimada de pertenecer a la segunda población, p_i, es pequeña, lo que indica que esta observación está mal explicada por el modelo.

• Si $y_i = 0$, y la observación pertenece a la primera población, sólo interviene el segundo término de la desviación d_i = -2 $\log(1 - p_i)$. La desviación será grande si p_i es grande, lo que indica que la probabilidad de pertenecer a la verdadera población es pequeña y el modelo ajusta mal dicho dato.

Para maximizar la verosimilitud, expresando p_i en función de los parámetros de interés, β ,en (5.6) obtenemos la función soporte:

$$L(\beta) = \sum_{i=1}^{n} y_i \mathbf{x}_i' \beta - \sum_{i=1}^{n} \log\left(1 + e^{\mathbf{x}_i'\beta}\right)$$

que derivaremos para obtener los estimadores MV. Escribiendo el resultado como vector columna:

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{n} y_i \mathbf{x}_i - \sum_{i=1}^{n} \mathbf{x}_i \left(\frac{e^{\mathbf{x}_i'\beta}}{1 + e^{\mathbf{x}_i'\beta}}\right) \tag{5.9}$$

e igualando este vector a cero y llamando $\widehat{\beta}$ a los parámetros que satisfacen el sistema de ecuaciones:

$$\sum_{i=1}^{n} y_i \mathbf{x}_i = \sum \mathbf{x}_i \left(\frac{1}{1 + e^{-\mathbf{x}_i'\widehat{\beta}}}\right) = \sum \widehat{y}_i \mathbf{x}_i \tag{5.10}$$

Estas ecuaciones establecen que el producto de los valores observados por las variables explicativas debe ser igual al de los valores previstos. También, que los residuos del modelo, $e_i = y_i - \widehat{y}_i$, deben ser ortogonales a las variables $\mathbf{x}$. Esta condición es análoga a la obtenida en el modelo de regresión estándar, pero ahora el sistema (5.10) resultante no es lineal en los parámetros $\widehat{\beta}$. Para obtener el valor $\widehat{\beta}_{MV}$ que maximiza la verosimilitud acudiremos a un algoritmo tipo Newton-Raphson. Desarrollando el vector $(\partial L(\beta)\ /\partial\beta)$ alrededor de un punto β_a, se tiene

$$\frac{\partial L(\beta)}{\partial \beta} = \frac{\partial L\left(\beta_\alpha\right)}{\partial \beta} + \frac{\partial^2 L\left(\beta_\alpha\right)}{\partial \beta \partial \beta'}\left(\beta - \beta_a\right)$$

para que el punto β_a corresponda al máximo de verosimilitud su primera derivada debe anularse. Imponiendo la condición $\frac{\partial L(\beta_a)}{\partial \beta} = 0$, se obtiene:

$$\beta_a = \widehat{\beta} + \left(-\frac{\partial^2 L\left(\beta_a\right)}{\partial \beta \partial \beta'}\right)^{-1} \left(\frac{\partial L(\beta)}{\partial \beta}\right) \tag{5.11}$$

que expresa cómo obtener el punto máximo β_a, a partir de un punto próximo cualquiera β. La ecuación depende de la matriz de segundas derivadas, que, en el óptimo, es la inversa de la matriz de varianzas y covarianzas asintótica de los estimadores MV . Para obtener su expresión, derivando por segunda vez en (5.9), se obtiene:

$$\widehat{\mathbf{M}}^{-1} = \left(-\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta'} \right) = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \omega_i \tag{5.12}$$

donde los coeficientes ω_i están dados por:

$$\omega_i = \frac{e^{x_i'\beta}}{\left(1 + e^{x'/2}\right)^2} = p_i \left(1 - p_i\right)$$

Sustituyendo en (5.11) las expresiones (5.12) y (5.9) y evaluando las derivadas en un estimador inicial $\hat{\beta}$, se obtiene el siguiente método para obtener un nuevo valor del estimador, β_a, a partir del $\hat{\beta}$

$$\beta_a = \hat{\beta} + \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \hat{\omega}_i \right)^{-1} \left(\sum \mathbf{x}_i \left(y_i - \hat{p}_i \right) \right)$$

donde $\hat{p}_i$ y $\hat{\omega}_i$ se calculan con el valor $\hat{\beta}$. El algortimo puede escribirse como:

$$\beta_a = \hat{\beta} + \left(\mathbf{X}' \widehat{\mathbf{W}} \mathbf{X} \right)^{-1} \mathbf{X}' (\mathbf{Y} - \widehat{\mathbf{Y}}) \tag{5.13}$$

donde $\hat{\mathbf{W}}$ es una matriz diagonal con términos $\hat{p}_i \ (1 - \hat{p}_i)$ y $\widehat{\mathbf{Y}}$ el vector de valores esperados de $\widehat{\mathbf{Y}}$. La matriz de varianzas y covarianzas de los estimadores así obtenidos es aproximadamente, según (5.13), $(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}$. Observemos que la ecuación (5.13) indica que debemos modificar el estimador si los residuos no son ortogonales a las variables explicativas, es decir si $\mathbf{X}'(\mathbf{Y} - \hat{\mathbf{Y}}) \neq 0$. La modificación del estimador depende de esta diferencia y se reparte entre los componentes de $\hat{\beta}$ en función de su matriz de varianzas y covarianzas estimada.

La forma habitual de implementar este método es el siguiente algoritmo iterativo que proporciona en convergencia el estimador MV de β.

1. Fijar un valor arbitrario inicial, $\hat{\beta}_1$, para los parámetros y obtener el vector $\hat{\mathbf{Y}}_1$ para dicho valor en el modelo *logit*. Por ejemplo, si $\hat{\beta}_1 = 0$,

$$\hat{y}_i = \hat{p}_i = \frac{1}{1 + e^{-0}} = \frac{1}{2}$$

y el vector $\hat{\mathbf{Y}}$ tiene todas sus componentes iguales a 1/2.

2. Definir una variable auxiliar z_i de residuos estandarizados por:

$$z_i = \frac{y_i - \widehat{y}_i}{\sqrt{\widehat{y}_i\,(1-\widehat{y}_i)}} = \frac{y_i - \widehat{p}_i}{\sqrt{\widehat{p}_i\,(1-\widehat{p}_i)}}$$

o vectorialmente:

$$\mathbf{Z} = \widehat{\mathbf{W}}^{-1/2}(\mathbf{Y} - \widehat{\mathbf{Y}})$$

donde $\widehat{\mathbf{W}}$ es una matriz diagonal con términos $\widehat{y}_i(1-\widehat{y}_i)$.

3. Estimar por mínimos cuadrados una regresión con variable dependiente $\mathbf{Z}$ y matriz de regresores $\mathbf{T} = \widehat{\mathbf{W}}^{1/2}\mathbf{X}$. Los parámetros estimados con esta regresión, $\widehat{b}_1$, vendrán dados por:

$$\begin{aligned}\widehat{\mathbf{b}}_1 &= (\mathbf{T}'\mathbf{T})^{-1}\,\mathbf{T}'\mathbf{Z} \\ &= \left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{Y} - \widehat{\mathbf{Y}})\end{aligned}$$

y, comparando con (5.13), vemos que $\mathbf{b}_1$ estima el incremento $\beta_a - \widehat{\beta}_1$ de los parámetros que nos acerca al máximo.

4. Obtener un nuevo estimador de los parámetros $\widehat{\beta}_2$ del modelo logístico mediante

$$\widehat{\beta}_2 = \widehat{\beta}_1 + \widehat{\mathbf{b}}_1$$

5. Tomar el valor estimado resultante de la etapa anterior, que en general llamaremos $\widehat{\beta}_h$, y sustituirlo en la ecuación del modelo logístico para obtener el vector de estimadores $\widehat{\mathbf{Y}}(\widehat{\beta}_h) = \widehat{\mathbf{Y}}_h$. Utilizando este vector $\widehat{\mathbf{Y}}_h$ construir la matriz $\widehat{\mathbf{W}}_h$ y las nuevas variables $\widehat{\mathbf{Z}}_h$ y $\widehat{\mathbf{T}}_h$

$$\begin{aligned}\mathbf{Z}_h &= \widehat{\mathbf{W}}_h^{-1/2}\left(\mathbf{Y} - \widehat{\mathbf{Y}}_h\right) \\ \mathbf{T}_h &= \widehat{\mathbf{W}}_h^{1/2}\mathbf{X}\end{aligned}$$

y volver a la etapa 2. El proceso se repite hasta obtener la convergencia ($\widehat{\beta}_{h+1} \simeq \widehat{\beta}_h$).

5.6 Contrastes

Si queremos contrastar si una variable o grupo de variables incluidas dentro de la ecuación es significativo, podemos construir un contraste de la razón de verosimilitudes comparando el máximo de la función de verosimilitud para el modelo con y sin estas variables. Supongamos que $\beta = (\beta_1 \beta_2)$, donde β_1 tiene dimensión $p - s$, y β_2 tiene dimensión s. Se desea contrastar si el vector de parámetros:

$$H_0 : \beta_2 = 0$$

frente a la alternativa

$$H_1 : \beta_2 \neq 0$$

El contrate de razon de verosimilitudes utiliza que $\lambda = 2L(H_1) - 2L(H_0)$, donde $L(H_1)$ es el máximo del soporte cuando estimamos los parámetros bajo H_1 y $L(H_0)$ es el máximo cuando estimamos los parámetros bajo H_0 es, si H_0 es cierta, una χ^2_s. Una manera equivalente de definir el contraste es llamar $D(H_0) = -2L(\hat{\beta}_1)$ a la desviación cuando el modelo se estima bajo H_0, es decir, suponiendo que $\hat{\beta}_2 = 0$, y $D(H_1) = -2L(\hat{\beta}_1 \hat{\beta}_2)$ a la desviación bajo H_1. La desviación será menor con el modelo con más parámetros (la verosimilitud será siempre mayor bajo H_1 y, si H_0 es cierta, la diferencia de desviaciones, que es el contraste de verosimilitudes

$$\chi^2_s = D\left(H_0\right) - D\left(H_1\right) = 2L\left(\hat{\beta}_1 \hat{\beta}_2\right) - 2L\left(\hat{\beta}_1\right)$$

se distribuye como una χ^2_s con s grados de libertad.

En particular este test puede aplicarse para comprobar si un parámetro es significativo y debe dejarse en el modelo. Sin embargo, es más habitual en estos casos comparar el parámetro estimado con su desviación típica. Los cocientes

$$w_j = \frac{\hat{\beta}_j}{s\left(\hat{\beta}_j\right)}$$

se denominan estadísticos de $Wald$ y en muestras grandes se distribuyen, si el verdadero valor del parámetro es cero, como una normal estándar.

Una medida global del ajuste es

$$R^2 = 1 - \frac{D(\hat{\beta})}{D_0} = 1 - \frac{L(\hat{\beta})}{L\left(\beta_0\right)}$$

donde el numerador es la desviación (verosimilitud en el máximo) para el modelo con parámetros estimados $\hat{\beta}$ y el denominador la desviación (verosimilitud) para el modelo que sólo incluye la constante β_0. Observemos que, en este último caso, la estimación de la probabilidad p_i es constante para todos los datos e igual a m/n siendo m el número de elementos en la muestra con la variable $y = 1$. Entonces, sustituyendo en (5.7) la desviación máxima que corresponde al modelo más simple posible con sólo β_0 que asigna la misma probabilidad a todos los datos, es

$$D_0 = -2L(\beta_0) = -2m \log m - 2(n-m)\log(n-m) + 2n \log n$$

Por otro lado, si el ajuste es perfecto, es decir todas las observaciones con $y = 1$ tienen $p_i = 1$ y las de $y = 0$ tienen $p_i = 0$, entonces, según (5.6) la desviación es cero y $L(\hat{\beta}) = 0$ y $R^2 = 1$. Por el contrario, si las variables explicativas no influyen nada la desviación con las variables explicativas será igual que sin ellas, $L(\hat{\beta}) = L(\beta_0)$ y $R^2 = 0$.

5.7 Ejemplos 1: MEDIFIS

Vamos a utilizar los datos de *MEDIFIS* para construir un modelo *logit* que clasifique una persona como hombre o mujer en función de sus medidas físicas. Si intentamos explicar la variables binaria, género, en función de todas las variables observadas obtenemos que el modelo es:

```
Datos <- read.table("DatosMEDIFIS.txt") # MEDIFIS
colnames(Datos) <- c("genero","esta", "peso", "pie",
                     "lonb", "anes", "dcra", "lrt")

# Ajuste de un modelo logístico.
glm(genero ~ ., data = Datos, family = "binomial")
```

```
##
## Call:  glm(formula = genero ~ ., family = "binomial", data = Datos)
##
## Coefficients:
## (Intercept)         esta         peso          pie         lonb         anes
##   -637.8802      -4.9068       0.1494      15.3614       6.0690       3.7729
##
## Degrees of Freedom: 26 Total (i.e. Null);  19 Residual
## Null Deviance:       37.1
## Residual Deviance: 8.764e-10     AIC: 16
```

$$\log\left(\frac{p_i}{1-p_i}\right) = -637.88+15.36pie+3.77aes-6.45dcr+13.7lrt-4.91est+6.07lbr+0.15pes$$

el modelo ajusta perfectamente los datos y no es posible calcular desviaciones típicas para los coeficientes. El modelo no es único. El problema es que tenemos sólo 27 observaciones y con las siete variables clasificamos fácilmente todas las observaciones. Sin embargo, el modelo obtenido puede ser muy malo para clasificar otras observaciones.

Vamos a contruir el modelo paso a paso. La variable con mayor coeficiente en la ecuación anterior es el pie con lo que comenzaremos con esta variable. Estimando el modelo estimado con RStudio,el modelo es:

```
# Ajuste de un modelo logístico.
modelo <- glm(genero ~ pie, data = Datos, family = "binomial")
```

```
##
## Call:  glm(formula = genero ~ pie, family = "binomial", data = Datos)
##
## Coefficients:
## (Intercept)          pie
##     -795.51        20.38
##
## Degrees of Freedom: 26 Total (i.e. Null);  25 Residual
## Null Deviance:       37.1
## Residual Deviance: 3.819     AIC: 7.819
```

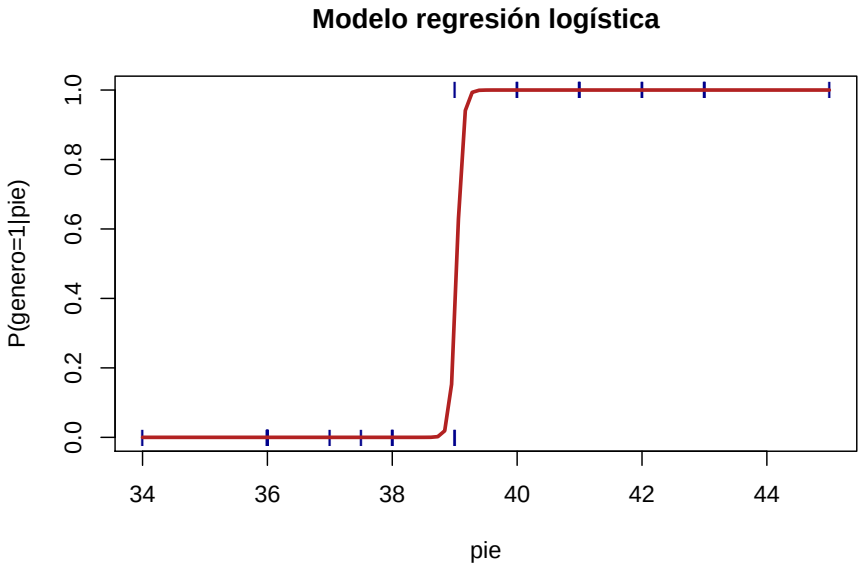
$$\log\frac{p_i}{1-p_i} = -795.51 + 20.38pie$$

Los dos parámetros están muy correlados y las desviaciones típicas son muy grandes. El valor inicial de la desviación es $D_0 = 37.1$. Después de estimar el modelo la desviación es $D = 3.8$. La diferencia entre desviaciones nos proporciona el contraste para ver si la variable *pie* es significativa. Esta diferencia es 33.27 que bajo la hipótesis de que el parámetro es cero será aproximadamente una distribución $Chi-cuadrado$ con 1 grado de libertad. El valor es tan grande que rechazamos la hipótesis a cualquier nivel de significación y concluimos que el *pie* es muy útil para discriminar.

5.7.1 Gráfico del modelo

Dado que un modelo logístico modela el logaritmo de $ODDs$, estas son las unidades en las que se devuelven las predicciones. Es necesario convertirlas de

nuevo en probabilidad mediante la función logit. En R, la función "predict()" puede devolver directamente las probabilidades en lugar de los *logODDs* si se indica el argumento type="response".



A la hora de evaluar la validez y calidad de un modelo de regresión logística, se analiza tanto el modelo en su conjunto como los predictores que lo forman. Se considera que el modelo es útil si es capaz de mostrar una mejora explicando las observaciones respecto al modelo nulo (sin predictores). El test *Likelihood* ratio calcula la significancia de la diferencia de residuos entre el modelo de interés y el modelo nulo. El estadístico sigue una distribución $Chi-cuadrado$ con grados de libertad equivalentes a la diferencia de grados de libertad de los dos modelos.

```
# Esto mismo se puede ver en R con la función "anova"

anova(modelo, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: genero
##
## Terms added sequentially (first to last)
##
##
```

Cuadro 5.1: Matriz de confusión de los datos de Medifis con el método de Regresión Logística aplicando la función a los mismos datos de entrenamiento

	0	1
0	15	0
1	1	11

```
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                    26     37.096
## pie   1   33.277        25      3.819 7.993e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Es interesante que, en este caso, el estadístico de *Wald* lleva a un resultado distinto. Como los parámetros están muy correlados, la desviación típica del coeficiente del *pie* es (en summary(modelo) dice eso) 6888.63, y el estadístico de *Wald* es $20.38/6888.63 = 0.003$ que no es significativo.

```
### Comparación de las predicciones con las observaciones

# Para este estudio se va a emplear un umbral de 0.5. Si la
# probabilidad predicha del género de la persona es superior
# a 0.5 se asigna al nivel 1 (hombre), si es menor se asigna
# al nivel 0 (mujer).

predicciones <- ifelse(test = modelo$fitted.values > 0.5,
                       yes = 1, no = 0)
t <- table(modelo$model$genero, predicciones,
           dnn = c("Clase real", "Clase predicha"))
```

Si aplicamos esta ecuación para clasificar los datos muestrales obtenemos un porcentaje de éxitos del 96%. Sólo una observación se clasifica mal como indica el Cuadro 5.1.

Si comparemos estos resultados con los del capítulo anterior (ejemplo de la sección 3.7) son bastante consistentes porque allí ya observamos que la variable más importante era el pié. El porcentaje de clasificación de la función lineal discriminante era del 100% que disminuía al 85% con validación cruzada. En el modelo logístico con sólo una variable hemos obtenido el 96% de éxito. Con validación cruzada este valor disminuye al 89% como indica el Cuadro 5.2.

Pero disminuye mucho menos que en el ejemplo de la sección 3.7 debido a la economía de parámetros que hace que se produzca menos sobreajuste.

Vamos a intentar introducir una variable adicional en el modelo logístico anterior que contiene sólo el *pie*. Introducimos la estatura, y el modelo estimado es

Cuadro 5.2: Matriz de confusión de los datos de Medifis aplicando LOOCV con el método de Regresión Logística

	0	1
0	13	2
1	1	11

```
# Ajuste de un modelo logístico.
modelo2 <- glm(genero ~ pie + esta, data = Datos, family = "binomial")
```

```
##
## Call:  glm(formula = genero ~ pie + esta, family = "binomial", data = Datos)
##
## Coefficients:
## (Intercept)          pie         esta
##   -803.3105      21.2938      -0.1693
##
## Degrees of Freedom: 26 Total (i.e. Null);  24 Residual
## Null Deviance:       37.1
## Residual Deviance: 3.709     AIC: 9.709
```

$$\log \frac{p_i}{1-p_i} = -803.31 + 21.29pie - 0.169est$$

```
summary(modelo2)
```

```
##
## Call:
## glm(formula = genero ~ pie + esta, family = "binomial", data = Datos)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.064   0.000   0.000   0.000   1.446
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.033e+02  2.588e+05  -0.003    0.998
## pie          2.129e+01  6.635e+03   0.003    0.997
## esta        -1.693e-01  5.243e-01  -0.323    0.747
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 37.0959  on 26  degrees of freedom
## Residual deviance:  3.7087  on 24  degrees of freedom
## AIC: 9.7087
##
## Number of Fisher Scoring iterations: 22
```

con desviaciones típicas $(2.588e+05)$, (6635) y (0.5243) . El coeficiente de estatura es -0.1693 con error estándar 0.5243 dando lugar a un estadístico de Wald de 0.3, con lo que concluimos que este coeficiente no es significativo. La desviación de este modelo es 3.709. Por tanto la reducción en desviación debida a la variable estatura con respecto al modelo que sólo incluye el pié es sólo de $3.80 - 3.71 = 0.09$, que no es significativa comparada con una $Chi - cuadrado$ con 1 grado de libertad.

```
anova(modelo2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: genero
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                    26     37.096
## pie   1   33.277        25      3.819 7.993e-09 ***
## esta  1    0.110        24      3.709    0.7397
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Este resultado es previsible ya que el modelo con las siete variables tiene una desviación de cero y el que contiene sólo el pie una desviación de $D = 3.8$: el contraste de que las seis variables adicionales no influyen lleva, en consecuencia, a un valor del estadístico de 3.8, y este valor en la hipótesis de que las variable no influyen debe provenir de una χ^2_6 con seis grados de libertad, lo que concuerda con lo observado, y debemos concluir que ninguna de las variables adicionales influye. La conclusión es pues que únicamente con el pie podemos clasificar estos datos con poco error.

5.8 Ejemplo 3: MUNDODES

Vamos a utilizar los datos de *MUNDODES* para ver cuáles son las variables que clasifican mejor a un país como perteneciente al continente africano. La función *logit* estimada es:

```
Datos <- read.table("DatosMUNDODES.txt")[,-8] #MUNDODES
colnames(Datos) <- c("Pais","TasaNat", "TasaMort",
                     "MortInf", "EspViH", "EspViM", "LPNB")
Datos[,7] <- log(Datos$LPNB)


# Como el continente africano esta codificado en los grupos como 5,
# vamos a transormar nuestra variable de respuesta a binaria,
# 0 = si no pertenecen al continente africano, 1 = si pertenecen.

Datos[,1] <- ifelse(Datos[,1] == 5, 1, 0)

modelo3 <- glm(Pais ~ ., data = Datos , family =binomial)
```

```
##
## Call:  glm(formula = Pais ~ ., family = binomial, data = Datos)
##
## Coefficients:
## (Intercept)      TasaNat     TasaMort      MortInf       EspViH       EspViM
##    15.58619      0.18632     -0.14202     -0.03344     -0.47293      0.13007      0
##
## Degrees of Freedom: 90 Total (i.e. Null);  84 Residual
## Null Deviance:       110.7
## Residual Deviance: 41.41     AIC: 55.41
```

$$\log \frac{p_i}{1-p_i} = 15.59 + .19tn - .14tm - .033mi - .47eh + .13em + .05lpnb$$

```
summary(modelo3)
```

```
##
## Call:
## glm(formula = Pais ~ ., family = binomial, data = Datos)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.80759  -0.27907  -0.07204   0.22692   2.49968
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 15.58619   14.18068   1.099  0.27172
## TasaNat      0.18632    0.07106   2.622  0.00874 **
## TasaMort    -0.14202    0.23652  -0.600  0.54821
## MortInf     -0.03344    0.02325  -1.439  0.15028
## EspViH      -0.47293    0.25358  -1.865  0.06218 .
## EspViM       0.13007    0.19294   0.674  0.50022
## LPNB         0.05123    0.46757   0.110  0.91275
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 110.664  on 90  degrees of freedom
## Residual deviance:  41.414  on 84  degrees of freedom
## AIC: 55.414
##
## Number of Fisher Scoring iterations: 8
```

Las variables tn, mi y eh son significativas, con un cociente entre la estimación del parámetro y la desviación típica de 6.8, 2.09 y 2.5 respectivamente. La desviación inicial es $D(\beta_0) = -2(27log27+64log64-91log91) = 110.66$. Por otro lado, la desviación del modelo estimado es $-2L(\hat{\beta}) = 41.41$ y la diferencia entre estas dos cantidades proporciona el contraste de que las variables no influyen. Si esto es cierto la diferencia, 69.25 será una $Chi - cuadrado$ con 6 grados de libertad. Como este valor es muy grande rechazamos esta hipótesis y admitimos que las variables influyen.

Además del valor de las estimaciones de los coeficientes parciales de correlación del modelo, es conveniente calcular sus correspondientes intervalos de confianza. En el caso de regresión logística, estos intervalos suelen calcularse empleando el método de profile likelihood (en R es el método por defecto si se tiene instalado el paquete MASS).

```
confint(object = modelo3, level = 0.95 )
```

```
##                    2.5 %       97.5 %
## (Intercept) -10.69149351 46.319507768
## TasaNat       0.06171243  0.350455511
## TasaMort     -0.70472959  0.192513553
## MortInf      -0.08458083  0.009651212
## EspViH       -1.04266582 -0.016809396
## EspViM       -0.27082561  0.527253140
## LPNB         -0.87649333  1.006486725
```

5.8.1 Evaluación del modelo

```
anova(modelo3, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Pais
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                        90    110.664
## TasaNat   1   62.402        89     48.262  2.8e-15 ***
## TasaMort  1    0.546        88     47.716  0.45994
## MortInf   1    0.328        87     47.388  0.56669
## EspViH    1    5.459        86     41.929  0.01947 *
## EspViM    1    0.503        85     41.426  0.47820
## LPNB      1    0.012        84     41.414  0.91266
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El modelo en conjunto sí es significativo y, acorde a los p-values mostrados en el summary(), también es significativa la contribución al modelo de ambos predictores.

El pseudo coeficiente de determinación es

$$R^2 = 1 - \frac{41.41}{110.66} = .63$$

5.8.2 Comparación de las predicciones con las observaciones

El Cuadro de clasificación con este modelo es

```
##    predicciones
##      0  1
##   0 61  3
##   1  5 22
```

que supone una proporción de éxitos de 83/91, o de 91%.

Usando validación cruzada:

```
##
##       0  1
##   0 59  5
##   1  6 21
```

5.9 Diagnosis

Los residuos del modelo logit (que a veces se denominan residuos de Pearson) se definen por:

$$e_i = \frac{y_i - \widehat{p_i}}{\sqrt{\widehat{p}_i\left(1-\widehat{p}_i\right)}}$$

y, si el modelo es correcto, serán variables de media cero y varianza unidad que pueden servirnos para hacer la diagnosis del modelo. El estadístico $\chi^2 = \sum_{i=1}^{n} e_i^2$ permite realizar un contraste global de la bondad del ajuste. Si el modelo es adecuado se distribuye asintóticamente como una χ^2 con $gl = n-p-1$, donde $p+1$ es el número de parámetros en el modelo.

En lugar de los residuos de Pearson se utiliza mucho las desviaciones de las observaciones o pseudoresiduos, definidas en (5.8) por $d_i = -2(y_i \log \widehat{p}_i + (1-y_i)\log(1-\widehat{p}_i))$, que aparece, de manera natural, en la maximización de la función de verosimilitud.

Podemos hacer un contraste de razón de verosimilitudes de la bondad del modelo como sigue: la hipótesis nula será que el modelo es adecuado, es decir, las probabilidades pueden calcularse con el modelo logístico con $p+1$ parámetros. La hipótesis alternativa será que el modelo no es adecuado y las n probabilidades son libres (supuesto que las x son distintas). Entonces, la desviación bajo H_0 es $D(H_0)$, mientras que bajo H_1 cada observación queda perfectamente clasificada dando $p_i = 0$ si pertenece a la primera población y $p_i = 1$ si pertenece a la segunda, y la desviación es cero porque todas las observaciones se clasifican sin error. El contraste de la razón de verosimilitudes se reduce al estadístico desviación global:

$$D\left(H_0\right) = -2\sum\left(y_i \log \widehat{p}_i + (1-y_i)\log\left(1-\widehat{p}_i\right)\right)$$

que, si el modelo es correcto, será también asintóticamente una χ^2 con $n-p-1$ grados de libertad.

5.10 El Modelo Multilogit

El modelo logit puede generalizarse para más de dos poblaciones, es decir, para variables respuesta cualitativas con más de dos niveles posibles. Supongamos G

poblaciones, entonces, llamando p_{ig} a la probabilidad de que la observación i pertenezca a la clase g, podemos escribir:

$$p_{ig} = \frac{e^{\beta_{0g}+\beta'_{1g}x_i}}{1+\sum_{j=1}^{G-1} e^{-\beta_{0j}-\beta'_{1j}x_i}} \quad j = 1, \ldots, G-1$$

y

$$p_{iG} = \frac{1}{1+\sum_{j=1}^{G-1} e^{-\beta_{0j}-\beta'_{1j}\mathbf{x}_i}}$$

con lo que automáticamente garantizamos que $\sum_{g=1}^{G} p_{ig} = 1$. Diremos que las probabilidades p_{ig} satisfacen una distribución logística multivariante. La comparación entre dos categorías se hace de la forma habitual

$$\frac{p_{ig}}{p_{ij}} = \frac{e^{\beta_{0g}+\beta'_{1g}\mathbf{x}_i}}{e^{\beta_{0j}+\beta'_{1j}\mathbf{x}_i}} = e^{(\beta_{0g}-\beta_{0j})} e^{\left(\beta'_{1g}-\beta'_{1j}\right)\mathbf{x}_i}$$

Esta ecuación indica que las probabilidades relativas entre dos alternativas no dependen del resto. Esa hipótesis puede generalizarse (véase Maddala (1983)).

La estimación y contrastes de esos modelos son extensiones directas de los *logit* ya estudiados y no entraremos en los detalles que el lector interesado puede encontrar en Fox (1984).

Capítulo 6

K vecinos más próximos (K-NN)

6.1 Introducción

Este es el último método de clasificación que veremos, es un procedimiendo simple y que ha dado buenos resultados con poblaciones no normales. Como el resto de los métodos de clasificación que hemos visto, se basa en calcular la probabilidad a posteriori, pero a diferencia de los demás métodos, K-NN es un metodo no paramétrico, esto significa que nuestro clasificador dependerá solamente de los datos o muestra que obtengamos sin conocer nada de la población y no hay que estimar ningún paramétro a partir de los datos.

6.2 Procedimiento operativo

El procedimiento es el siguiente:

(1) Definir una medida de distancia entre puntos, habitualmente la distancia euclidea usual o la distancia de Mahalanobis.

(2) Calcular las distancias del punto a clasificar, $\mathbf{x_0}$, a todos los puntos de la muestra.

(3) Seleccionar los K puntos muestrales más próximos al que pretendemos clasificar. Calcular la proporción de estos K puntos que pertenece a cada una de las poblaciones.

(4) Clasificar el punto $\mathbf{x_0}$ en la población con mayor frecuencia de puntos entre los K. Este método se conoce como K - vecinos más próximos (K-Nearest Neighbors).

En el caso particular de $K = 1$ el método consiste en asignarle a la población al que pertenece el elemento más próximo. Un problema clave de este método es claramente la selección de K. Una práctica habitual es tomar $K = \sqrt{n_g}$ donde n_g es un tamaño de grupo promedio. Otra posibilidad es probar con distintos valores de K, aplicarselo a los puntos de la muestra cuya clasificación es conocida, obtener el error de clasificación en función de K y escoger aquel valor de K que conduzca al menor error observado. En la sección 6.4.1 estudiamos el efecto de la elección de K.

Al principio dijimos que se basa en probabilidades a posteriori, esto se ve cuando definimos la probabilidad a posteriori como :

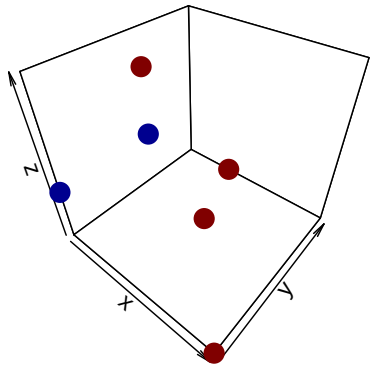
$$\Pr\left(Y = j \mid X = x_0\right) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I\left(y_i = j\right)$$

Dónde ese N_0 que aparece en la sumatoria son los K datos mas cercanos (de todas las poblaciones)

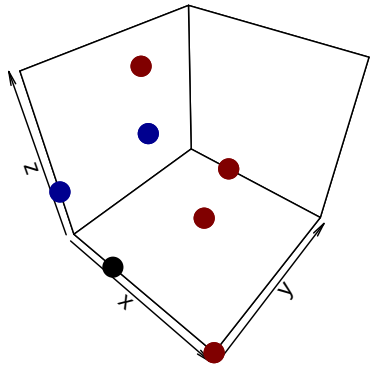
Para aclarar un poco como funciona este método veremos el siguiente ejemplo y haremos paso por paso el procedimiento operativo.

Ejemplo 1: La siguiente tabla proporciona un conjunto de datos de entrenamiento que contiene seis observaciones, tres predictores y una variable de respuesta cualitativa.

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Rojo
2	2	0	0	Rojo
3	0	1	3	Rojo
4	0	1	2	Azul
5	−1	0	1	Azul
6	1	1	1	Rojo



Supongamos que deseamos usar este conjunto de datos para hacer una predicción para Y cuando $X_1 = X_2 = X_3 = 0$ usando K vecinos más cercanos.



Calcularemos la distancia euclídea entre cada observación y el punto de prueba, $X_1 = X_2 = X_3 = 0$.

Obs.	*Distancias*	*Y*
1	3	Rojo
2	2	Rojo
3	3, 1623	Rojo
4	2, 2361	Azul
5	1, 442	Azul
6	1, 7321	Rojo

Los ordenaremos de menor a mayor para facilitar al momento de buscar los más cercanos.

Obs.	*Distancias*	*Y*
5	1, 442	Azul
6	1, 7321	Rojo
2	2	Rojo
4	2, 2361	Azul
1	3	Rojo
3	3, 1623	Rojo

Si elegimos $K = 1$ los puntos muestrales más próximos al que pretendemos clasificar el punto de prueba, $X_1 = X_2 = X_3 = 0$, la observación 5 es la más cercana, por lo tanto lo clasificaremos como Y = Azul.

Suponga que elegimos $K = 3$. Entonces K-NN identificará primero las tres observaciones que están más cerca del punto de prueba.

Obs.	*Distancias*	*Y*
5	1, 442	Azul
6	1, 7321	Rojo
2	2	Rojo

Consiste en dos puntos Rojos y un punto Azul, lo que da como resultado probabilidades estimadas de 2/3 para la clase Rojo y 1/3 para la clase Azul. Por lo tanto, K-NN predecirá que el punto de prueba pertenece a la clase Rojo.

Por esto mismo, es muy importante no tomarse a la ligera la elección del valor de K ya que puede cambiar el resultado de la predicción.

6.3 Ejemplo: MEDIFIS

Vamos a utilizar los datos de $MEDIFIS$ para clasificar personas por su género conocidas las medidas físicas de la variables.

```
# Cargamos los datos
Datos <- read.table("DatosMEDIFIS.txt") # MEDIFIS

# Cambiamos el nombre de las columnas
colnames(Datos) <- c("genero","esta", "peso", "pie",
                     "lonb", "anches", "diamcra", "lrt")
head(Datos)
```

```
##   genero esta peso pie lonb anches diamcra lrt
## 1      0  159   49  36   68   42.0      57  40
## 2      1  164   62  39   73   44.0      55  44
## 3      0  172   65  38   75   48.0      58  44
## 4      0  167   52  37   73   41.5      58  44
## 5      0  164   51  36   71   44.5      54  40
## 6      0  161   67  38   71   44.0      56  42
```

6.4 Implementación del método

```
# Crearemos una función que se llamará "clasificador" y seguiremos
# el procedimiento dicho al principio.

clasificador <- function(x, Datos, k) {
  #calcular distancia entre x y cada uno de las filas de Datos
  distancias = apply(Datos[,-1], 1, function(fila)
  {
    dist(rbind(x, fila))
  })
  names(distancias) = Datos[,1]

  #Ordenar las distancias

  #Seleccionar las k distancias menores
  tabla = table(names(sort(distancias)[1:k]))

  #Averiguar la frecuencia de los rotulos de las k distancias mas pequeñas

  #Reportar como clase el rotulo mas frecuenta
  names(tabla[which.max(tabla)])
}
```

Ahora para dar una noción más real de como funciona el método, usaremos LOOCV en los datos de MEDIFIS, dejando una fila sin usar de entrenamiento y la usaremos luego como una observación nueva a ser clasificada.

Cuadro 6.1: Matriz de confusión de los datos de Medifis aplicando LOOCV con el método de K-NN

	0	1
0	13	2
1	2	10

El Cuadro 6.1 muestra que aplicando validación cruzada se obtiene una proporción de aciertos de 23/27=0.852. Las observaciones mal clasificadas son las 2, 3, 22, y 25. Vemos que el método de validación cruzada da una idea más realista de la eficacia del procedimiento de clasificación.

Algo muy positivo que tiene K-NN, que no tienen los clasificadores lineales, es la flexibilidad de su limite de decisión, puede formar regiones de clasificación de lo mas complejos.

Por ejemplo, supongamos que tenemos dos conjuntos de datos como los presentados en la Figura 6.1, que representa dos clases, una marcada por los puntos rojos y otra marcada por los puntos negros. Apliquemos los clasificadores lineal, logit y knn y comparemos los resultados.

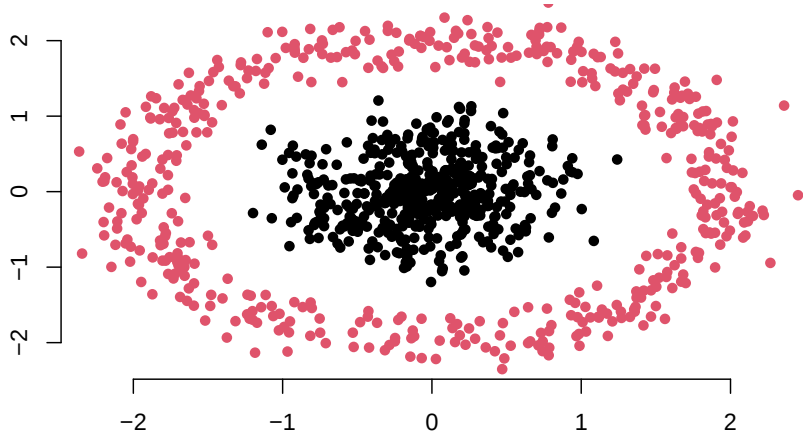


Figura 6.1: Diagrama de dispersión de datos simulados formando dos grupos con un limite de decisión no lineal

A simple vista vemos que las dos clases están separados por algún círculo y vemos la Figura 6.2 vemos que K-NN es el que mejor se acerca a un limite de decisión óptimo.

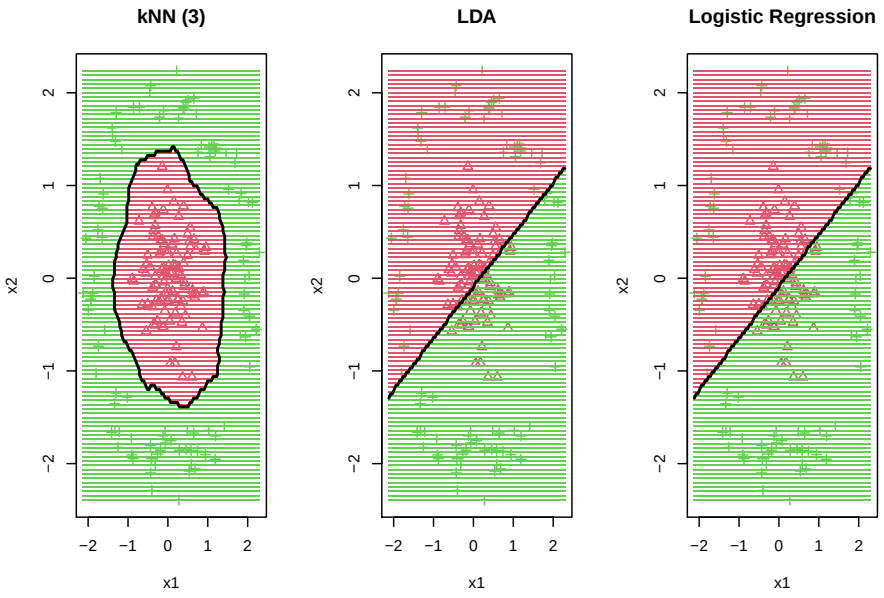


Figura 6.2: Limites de decisión de los clasificadores Izquierda: K-NN, Centro: LDA, Derecha: LOGIT

6.4.1 Efectos en la elección del K.

En esta sección vamos a estudiar el efecto de la elección del parámetro K en K-NN en el problema de clasificación. Mostraremos los cambios en las tasas de error y en los límites de decisión.

Generamos datos sintéticos en 5 escenarios distintos. Utilizaremos los escenarios 1, 3, 4, 5, 6 que serán explicados con detalle en el capítulo siguiente. En cada escenario se generaron 200 datos de entrenamiento (100 en cada categoría) y 400 de prueba (200 en cada categoría).

Esta simulación consiste en crear 5 conjuntos de datos, elegir un valor de K y calcular la tasa de error para ese valor del parámetro.

Al igual que en los métodos mencionados en los capítulos anteriores, no existe una fuerte relación entre la tasa de error de entrenamiento y la tasa de error de prueba. Con K = 1, la tasa de error de entrenamiento KNN es 0, pero la tasa de

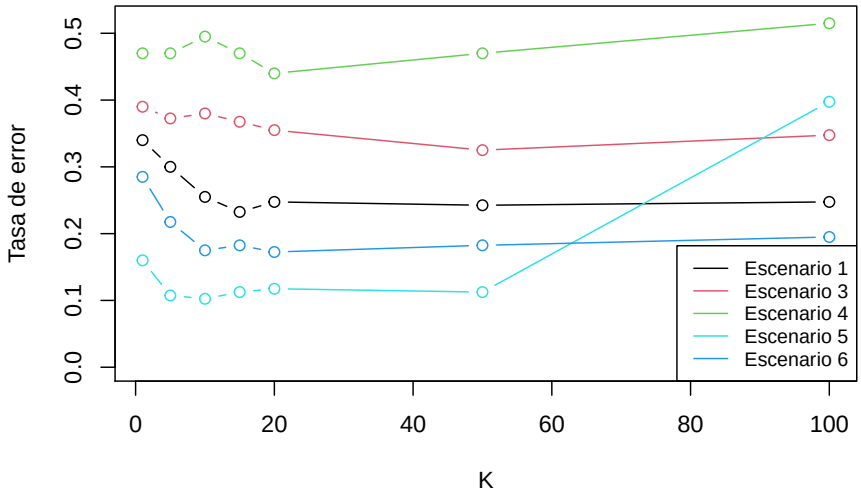


Figura 6.3: Tasa de error de prueba de K-NN en función de K en 5 conjuntos de datos simulados distintos.

error de prueba puede ser bastante alta. En general, a medida que utilizamos métodos de clasificación más flexibles, la tasa de error de entrenamiento disminuirá, pero es posible que la tasa de error de prueba no. En la Figura 6.3, hemos graficado la tasa de error de prueba de K-NN en función de K. A medida que aumenta K, el método se vuelve menos flexible. Notemos que el error de prueba, en algunos escenarios, exhibe una característica en forma de U, disminuyendo al principio (con un mínimo para K en el intervalo entre 5 y 20) antes de aumentar nuevamente cuando el método se vuelve menos flexible.

Observemos con un poco de detenimiento el caso del escenario 6 (frontera de decisión cuadrática 7.6). La Figura 6.4 muestra las verdaderas clases de los datos y la Figura 6.5 muestra los limites de decision en este caso para $K = 1, \ldots, 100$. Con $K = 1$, el límite de decisión es más flexible y encuentra patrones que no corresponden y con $K = 100$, el límite de decisión se va haciendo más lineal. En este conjunto de datos simulados, ni $K = 1$ ni $K = 100$ dan buenas predicciones: tienen tasas de error de prueba de 0.2850 y 0.1950, respectivamente, mientras que con $K = 10$ tenemos un error de prueba de 0.1750.

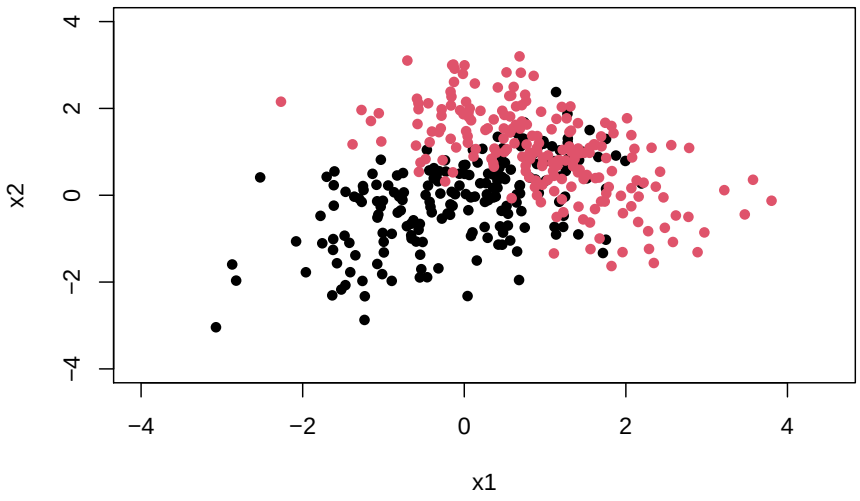


Figura 6.4: Diagrama de dispersión de los datos de prueba para el escenario 6 mostrados en la Figura 6.3 diferenciando por colores las dos clases distintas.

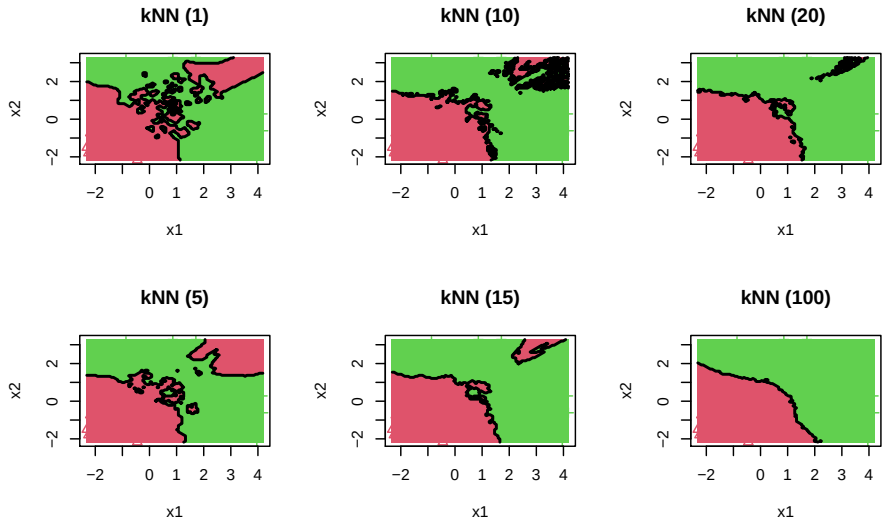


Figura 6.5: Comparación de los limites de decición de K-NN obtenido usando K=1, 5, 10, 15, 20 y 100 en otro de los 5 escenarios que mostramos en la Figura 6.3 cuyo límite de decisión es cuadrático.

Capítulo 7

Simulaciones

En este capítulo presentamos simulaciones de varios escenarios con el fin de comparar los métodos mencionados en los capítulos pasados y agregaremos uno más: K-NN aplicando validación cruzada como método para elegir el valor de K (lo llamaremos KNN.cv). Esto es, como explicamos en la seccion 2.4, se ajusta el método para varios valores de K y se elige aquel que arroja la menor tasa de error en los datos de entrenamiento. Consideramos los valores de K del 1 al 10 ya que nuestro conjunto de datos no es tan grande. El código con la función implementada para la elección del K junto con el código utilizado para las simulaciones que siguen se pueden encontrar en el apéndice B.

Realizamos siete escenarios, en cada uno simulamos 100 conjuntos de datos de entrenamiento. Luego ajustamos cada uno de los cinco métodos y calculamos la tasa de error, VP, FP, VN y FN en un conjunto de validación tamaño 200.

En cada escenario consideramos dos clases. Dentro de cada clase generamos 20 observaciones bivariadas (datos de entrenamiento) con dos predictoras (X_1 y X_2) de distintas distribuciones.

Para cada escenario mostramos diagramas de cajas y bigotes de la tasa de error obtenida y a modo de completitud mostramos tablas con las medidas ya mencionadas y agregamos la sensibilidad y la especificidad.

7.1 Escenario 1

Las observaciones dentro de cada clase se generan a partir de una distribución normal bivariada, la primera población con media $\mu_1 = (0, 0)$ y la segunda con media $\mu_2 = (2, 2)$, ambas clases tienen la misma matriz de varianza-covarianza

$$\mathbf{V} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

La Figura 7.1 muestra las tasas de error obtenidas en este escenario. Se puede observar que LDA se desempeñó bien en este entorno, como era de esperar, ya que este es el modelo asumido por LDA. QDA también clasificó peor que LDA, ya que ajusta un clasificador más flexible de lo necesario. Dado que la regresión logística asume un límite de decisión lineal, sus resultados fueron solo ligeramente inferiores a los de LDA.

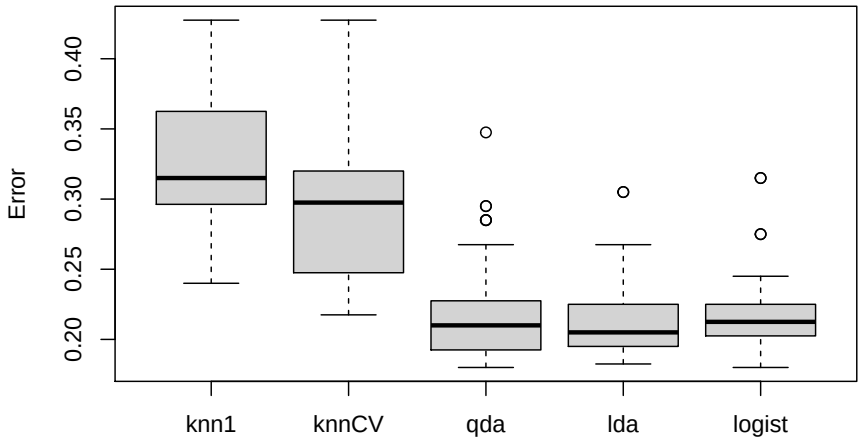


Figura 7.1: Diagrama de caja y bigotes de la tasa de error de cada método escenario 1

En el Cuadro 7.1 vemos que LDA y la regresión logística superan en todas las medidas al resto de los métodos, estando muy parejos entre ellos.

Cuadro 7.1: Medidas de todos los métodos escenario 1

	VP	VN	FP	FN	ACIER	ERROR	PREC	SENS	ESP
knn1	133.68	136.00	66.32	64.00	0.6742	0.3258	0.6684	0.6777	0.6727
knnCV	136.65	145.11	63.35	54.89	0.7044	0.2956	0.6832	0.7165	0.6979
qda	155.17	157.74	44.83	42.26	0.7823	0.2177	0.7758	0.7869	0.7819
lda	156.27	159.11	43.73	40.89	0.7884	0.2115	0.7814	0.7939	0.7860
logist	152.97	160.92	47.03	39.08	0.7847	0.2153	0.7648	0.7983	0.7759

7.2 Escenario 2

Los datos fueron simulados como en el escenario 1, excepto que dentro de cada clase, los dos predictores tienen una correlación de 0.5. La Figura 7.2 indica pocos cambios en el desempeño relativo de los métodos en comparación con el escenario anterior. Lo mismo se observa en el Cuadro 7.2, donde las medidas de evaluación de LDA y la regresión siguen muy parejas.

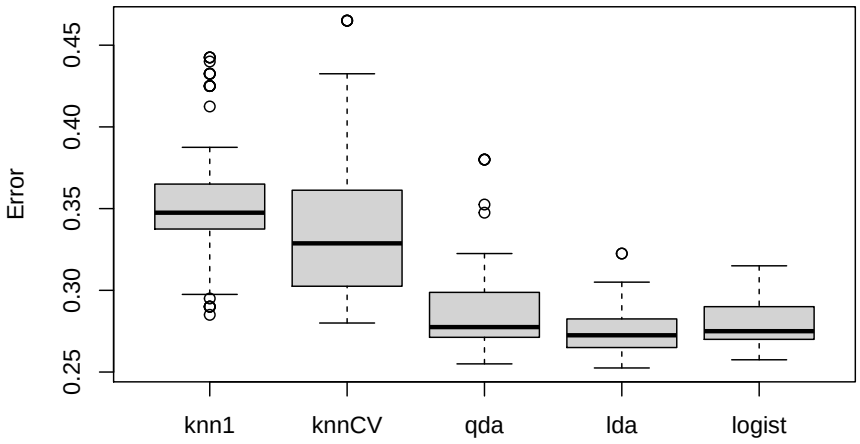


Figura 7.2: Diagrama de caja y bigotes de la tasa de error de cada método escenario 2

Cuadro 7.2: Medidas de todos los métodos escenario 2

	VP	VN	FP	FN	ACIER	ERROR	PREC	SENS	ESP
knn1	121.74	135.71	78.26	64.29	0.6436	0.3564	0.6087	0.6575	0.6340
knnCV	121.24	143.03	78.76	56.97	0.6607	0.3393	0.6062	0.6868	0.6434
qda	131.26	153.88	68.74	46.12	0.7128	0.2872	0.6563	0.7406	0.6937
lda	133.60	155.76	66.40	44.24	0.7234	0.2766	0.6680	0.7527	0.7026
logist	131.67	157.11	68.33	42.89	0.7220	0.2780	0.6584	0.7561	0.6987

Cuadro 7.3: Medidas de todos los métodos escenario 3

	VP	VN	FP	FN	ACIER	ERROR	PREC	SENS	ESP
knn1	122.12	116.18	77.88	83.82	0.5958	0.4042	0.6106	0.5923	0.6005
knnCV	128.41	128.12	71.59	71.88	0.6413	0.3587	0.6420	0.6433	0.6409
qda	141.04	127.73	58.96	72.27	0.6719	0.3281	0.7052	0.6660	0.6849
lda	138.73	137.28	61.27	62.72	0.6900	0.3100	0.6936	0.6897	0.6939
logist	137.43	138.82	62.57	61.18	0.6906	0.3094	0.6872	0.6933	0.6920

7.3 Escenario 3

Los datos provienen de una distribucion t-student bivariada con sus predictoras, X_1 y X_2, incorreladas dentro de cada clase, dónde las dos clases tienen los mismos grados de libertad (5) y distinto parametro de no centralidad (o de locación μ). Elegimos esta configuración debido a que los grados de libertad solo cambian el ancho y alto de la curva. La distribución t tiene una forma similar a la distribución normal, pero tiende a producir puntos más extremos, es decir, más puntos que están lejos de la media. En este contexto, el límite de decisión sigue siendo lineal y, por lo tanto, se ajusta al marco de regresión logística. La configuración viola los supuestos de LDA, ya que las observaciones no provienen de una distribución normal. La Figura 7.3 muestra que la regresión logística superó ligeramente al LDA, aunque ambos métodos fueron superiores a los otros enfoques. En particular, los resultados de QDA se deterioraron considerablemente como consecuencia de la no normalidad.

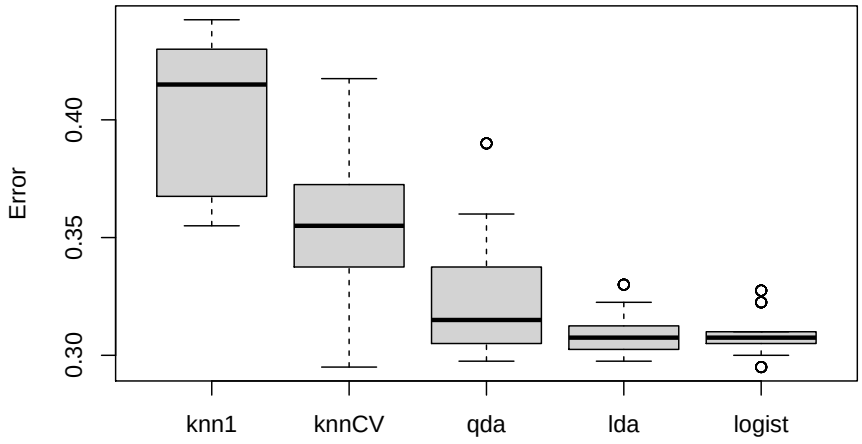


Figura 7.3: Diagrama de caja y bigotes de la tasa de error de cada método escenario 3

Cuadro 7.4: Medidas de todos los métodos escenario 4

	VP	VN	FP	FN	ACIER	ERROR	PREC	SENS	ESP
knn1	52.31	56.85	47.69	43.15	0.5458	0.4542	0.5231	0.5478	0.5450
knnCV	55.79	54.46	44.21	45.54	0.5512	0.4488	0.5579	0.5503	0.5526
qda	46.22	71.43	53.78	28.57	0.5883	0.4118	0.4622	0.6259	0.5708
lda	60.46	65.97	39.54	34.03	0.6322	0.3678	0.6046	0.6416	0.6261
logist	61.66	64.19	38.34	35.81	0.6292	0.3708	0.6166	0.6336	0.6273

7.4 Escenario 4

Los datos provienen de una distribucion Chi-cuadrado donde las dos clases tienen los mismos grados de libertad y distinto parámetro de no centralidad. La Figura 7.4 y el Cuadro 7.4 muestran que los resultados son muy parecidos al escenario 1, LDA es superior a la regresión logística a pesar de no cumplirse la hipótesis de normalidad. Este ejemplo y el anterior no revelan mayores cambios en el desempeño relativo de los métodos debido a la ausencia de normalidad.

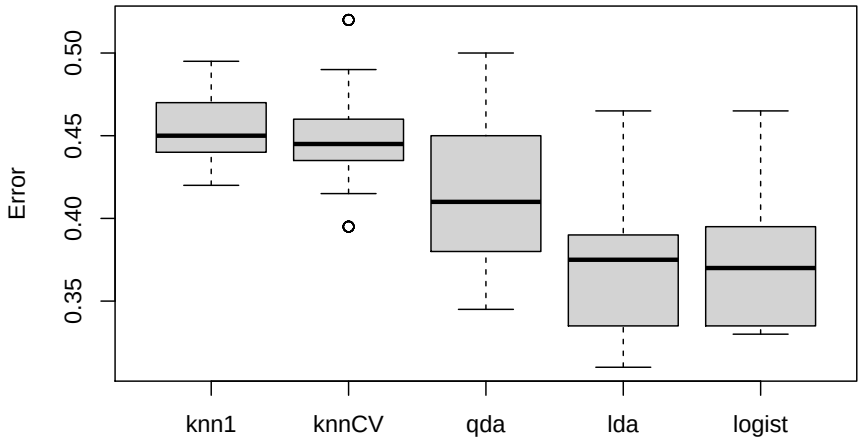


Figura 7.4: Diagrama de caja y bigotes de la tasa de error de cada método escenario 4

Cuadro 7.5: Medidas de todos los métodos escenario 5

	VP	VN	FP	FN	ACIER	ERROR	PREC	SENS	ESP
knn1	81.89	83.38	18.11	16.62	0.8264	0.1736	0.8189	0.8317	0.8255
knnCV	85.96	85.24	14.04	14.76	0.8560	0.1440	0.8596	0.8547	0.8608
qda	92.18	81.20	7.82	18.80	0.8669	0.1331	0.9218	0.8312	0.9128
lda	47.76	49.53	52.24	50.47	0.4864	0.5136	0.4776	0.4833	0.4891
logist	47.76	49.53	52.24	50.47	0.4864	0.5136	0.4776	0.4833	0.4891

7.5 Escenario 5

Dentro de cada clase, las observaciones se generaron a partir de una distribución normal con predictores no correlacionados. Luego, la mitad de los datos de una de las clases fueron trasladados hacia la izquierda mientras la otra mitad fue trasladada hacia la derecha, como se muestra en la Figura 7.5. La Figura 7.6 indica que K-NN aplicando validación cruzada y QDA se desempeñaron mejor, esto se debe a que el límite de decisión no es lineal.

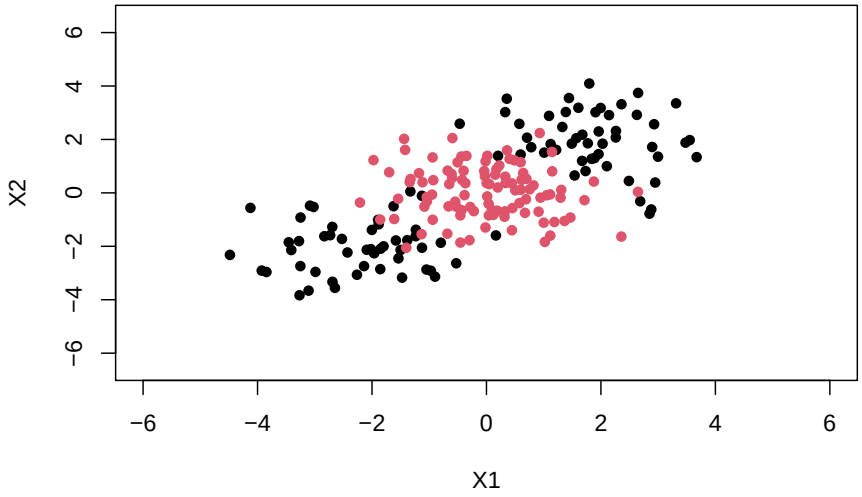


Figura 7.5: Gráfico de dispersión de los datos escenario 5

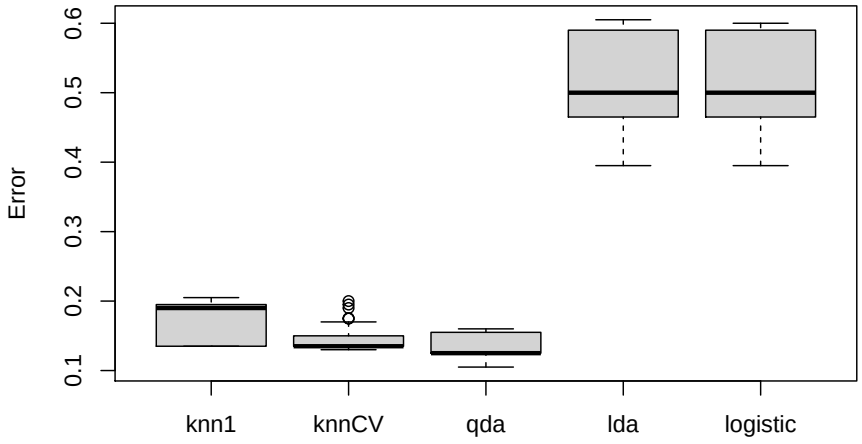


Figura 7.6: Diagrama de caja y bigotes de la tasa de error de cada método escenario 5

Cuadro 7.6: Medidas de todos los métodos escenario 6

	VP	VN	FP	FN	ACIER	ERROR	PREC	SENS	ESP
knn1	142.41	149.49	57.59	50.51	0.7298	0.2702	0.7120	0.7410	0.7221
knnCV	140.09	156.62	59.91	43.38	0.7418	0.2582	0.7004	0.7718	0.7226
qda	148.98	180.23	51.02	19.77	0.8230	0.1770	0.7449	0.8843	0.7798
lda	145.84	171.05	54.16	28.95	0.7922	0.2078	0.7292	0.8367	0.7601
logist	150.17	166.24	49.83	33.76	0.7910	0.2090	0.7508	0.8184	0.7697

7.6 Escenario 6

Los datos se generaron a partir de una distribución normal, con una correlación de 0.5 entre los predictores de la primera clase y una correlación de -0.5 entre los predictores de la segunda clase. Esta configuración corresponde al supuesto de QDA y resulta en límites de decisión cuadráticos. La Figura 7.7 y el Cuadro 7.6 muestran que QDA superó a todos los demás enfoques.

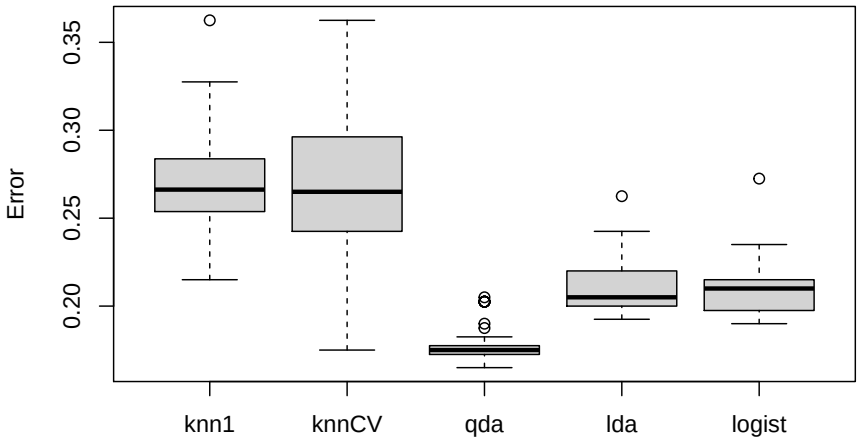


Figura 7.7: Diagrama de caja y bigotes de la tasa de error de cada método escenario 6

7.7 Escenario 7

En este escenario simulamos datos buscando un límite de decisión muy alejado de la linealidad. En la Figura 7.8 vemos una clase representada por los puntos rojos y otra clase representada por los puntos negros. Como resultado, incluso los límites de decisión cuadráticos de QDA no pudieron modelar adecuadamente los datos. La Figura 7.9 y el Cuadro 7.7 muestran que en este caso, los métodos lda, qda y reg. log. tienen muy mal desempeño, con resultados aún peores que se clasificáramos tirando una moneda al aire. Siendo QDA el que peores resultados da. Por otro lado, los métodos no paramétricos, KNN1 y KNN-CV arrojaron muy buenos resultados, esto se debe a la flexibilidad de estos métodos.

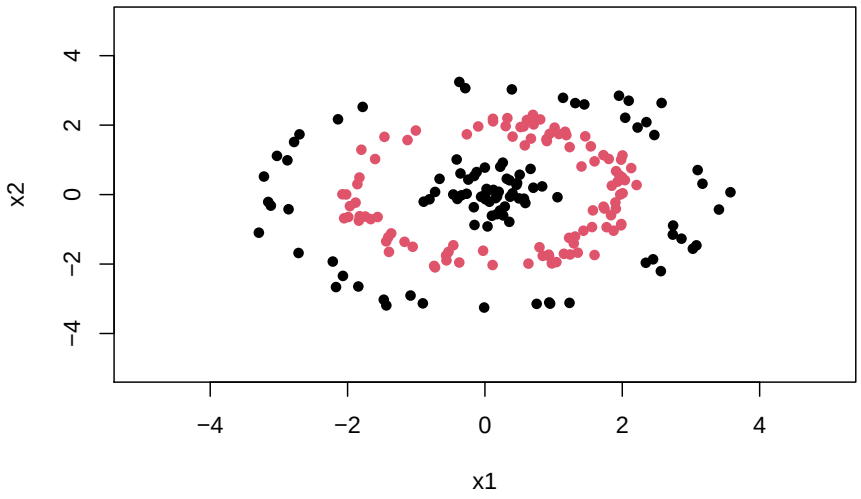


Figura 7.8: Gráfico de dispersión de los datos escenario 7

Cuadro 7.7: Medidas de todos los métodos escenario 7

	VP	VN	FP	FN	ACIER	ERROR	PREC	SENS	ESP
knn1	92.54	99.52	7.46	0.48	0.9603	0.0397	0.9254	0.9950	0.9308
knnCV	92.54	99.52	7.46	0.48	0.9603	0.0397	0.9254	0.9950	0.9308
qda	38.79	48.26	61.21	51.74	0.4353	0.5648	0.3879	0.4289	0.4405
lda	68.04	46.32	31.96	53.68	0.5718	0.4282	0.6804	0.5590	0.5922
logist	68.05	46.32	31.95	53.68	0.5718	0.4281	0.6805	0.5590	0.5922

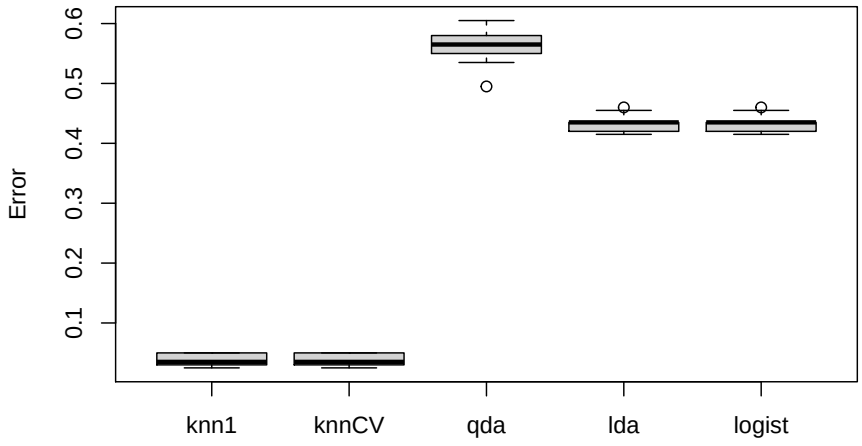


Figura 7.9: Diagrama de caja y bigotes de la tasa de error de cada método escenario 7

7.8 Conclusión

Estos siete ejemplos ilustran que ningún método dominará a los demás en todas las situaciones. Cuando los límites de decisión verdaderos son lineales, entonces los enfoques de regresión logística y LDA tenderán a funcionar bien. Cuando los límites son moderadamente no lineales, QDA puede dar mejores resultados. Finalmente, para límites de decisión mucho más complicados, un enfoque no paramétrico como KNN puede ser superior. Pero el nivel de suavidad para un enfoque no paramétrico debe elegirse con cuidado.

Capítulo 8

Aplicación a datos reales.

8.1 Introducción

En este capítulo utilizaremos los métodos estudiados en los capítulos anteriores para realizar un análisis en datos reales. Los datos provienen del proyecto de investigación "Trazabilidad de Contactos a través del Contexto Digital de Dispositivos Móviles" financiado por FONCyT a través de la Convocatoria Extraordinaria Ideas-Proyecto COVID 19. En el proyecto participan docentes investigadores y estudiantes de la Facultad de Ciencias Exactas, Físicas y Naturales (FCEFyN), la Facultad de Ciencias Médicas y la Facultad de Matemática, Astronomía, Física y Computación de la UNC y equipos de otras universidades del país. En el marco de este proyecto el Laboratorio de Comunicaciones Digitales de la FCEFyN de la UNC desarrolló una aplicación para celulares que permite registrar la señal Bluetooth entre dos dispositivos. Es parte de los objetivos del proyecto poder deducir la distancia entre dos celulares utilizando estos registros con el fin de poder realizar seguimiento de contactos cercanos. Este es un problema actual en el contexto de la pandemia COVID-19.

El problema de clasificación que abordaremos consiste en determinar si dos celulares están cerca o lejos dada la señal de Bluetooth que emiten o reciben. Cuando dos dispositivos están cerca es posible registrar la señal en un nivel de potencia atenuado conocido como indicador de intensidad de la señal recibida (RSSI). Para más detalle sobre la señal RSSI ver Rappaport et al. (1996).

Para abordar el problema seguiremos un trabajo desarrollado en el marco del proyecto que se encuentra enviado para su evaluación y posible publicación. En ese trabajo se utilizaron medidas resumen de la series de RSSI para clasificar la distancia entre los celulares de los experimentos en dos categorías: cerca (menos de 2 metros) o lejos (2 metros o más). Aquí reproduciremos parte del análisis de ese artículo pero utilizando una base más actual (más experimentos).

Cuadro 8.1: Cuadro de las primeras 6 filas de algunas medidas resumen correspoondiente a la señal RSSI.

mean	median	std	dis1	dis2	conteo	q1
-56.9294	-57	0.4020	0.1766	0.1176	4	-57
-64.6835	-65	0.5671	0.5018	0.4177	3	-65
-60.9494	-61	0.3162	0.1442	0.1013	3	-61
-62.5529	-63	0.5234	0.5049	0.4471	3	-63
-65.9726	-66	0.1644	0.0533	0.0274	2	-66
-48.8797	-50	7.6710	6.3567	6.3359	26	-53

Los datos con los que trabajaremos fueron obtenidos mediante experimentos controlados, consistentes en tomar los valores RSSI por una cantidad de tiempo determinada entre dos celulares. Cada experimento consistió en al menos dos teléfonos inteligentes en el mismo entorno que periódicamente emitieron y escanearon señales Bluetooth durante una ventana de observación, cada experimento se realizó mientras se miden cuidadosamente las distancias entre los dispositivos participantes (rangos de 0 a 4 metros). Es importante destacar que algunos experimentos se llevaron a cabo al aire libre (outdoor) y otros en ambientes cerrados (indoor).

La base con la que vamos a trabajar consiste en 2915 entradas. Cada una de las observaciones es un conjunto de medidas resumen correspondiente a la señal RSSI registrada entre dos dispositivos en un experimento dado. Además, para cada entrada se tiene la variable correspondiente al ambiente (environment) y la variable respuesta binarizada (cerca o lejos). Las medidas resumen consideradas son las siguientes: media, media recortada, mediana, primer y tercer cuartil, mínimo, máximo, desviación estándar, rango, rango intercuartil, distancia L_1 a la media ($\frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|$), la distancia L_1 a la mediana ($\frac{1}{n}\sum_{i=1}^{n}|x_i - \tilde{x}|$), curtosis, asimetría, y el recuento de diferentes valores que describen la serie. Éstas se utilizarán en sub grupos para entrenar los modelos de clasificación y estudiar cuales son más útiles para distinguir la variable respuesta. El Cuadro 8.1 muestra algunas de las variables estadísticas a utilizar para las primeras 6 entradas de la base.

La Figura 8.1 muestra el coeficiente de correlación de Pearson entre todas las variables de la base. Como se puede observar, alguna de ellas están fuertemente correlacionadas. Debido a este hecho, no consideraremos todas las combinaciones posibles, sino solo un subconjunto de ellas. De la figura, se pueden identificar 3 grupos dentro de los cuales las variables están fuertemente correlacionadas entre sí.

Como se resume en el Cuadro 8.2, los 3 grupos son los siguientes: medidas de posición (Grupo 1), medidas de dispersión (Grupo 2) y medidas de forma (Grupo 3). Las correlaciones dentro de los grupos para los grupos 1 y 2 son mayores

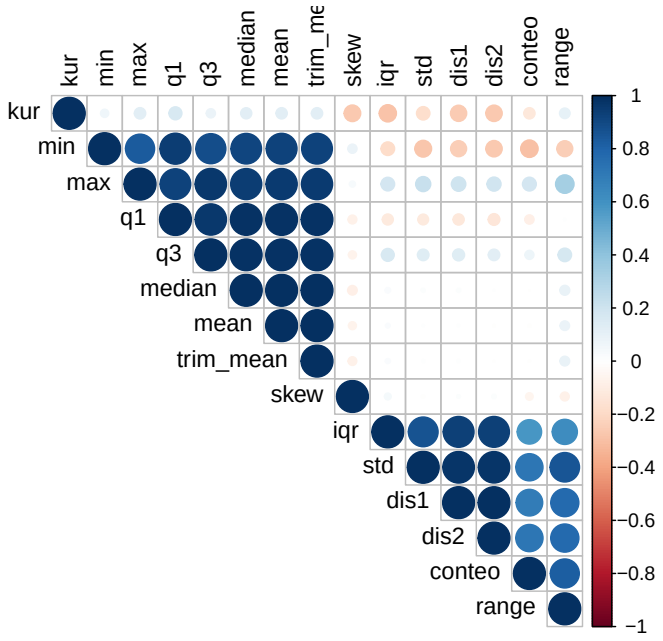


Figura 8.1: Correlación entre todas las variables.

Cuadro 8.2: Grupos de características.

Grupo 1	Grupo 2	Grupo 3
media (mean)	desviación estandar (std)	asimetría (skew)
media recortada (tmean)	rango (range)	curtosis (kur)
mediana (median)	rango intercuartil (iqr)	conteo (cnt)
primer cuartil (q1)	distancia L1 a la media (dis1)	
tercer cuartil (q3)	distancia L1 a la mediana (dis2)	
mínimo (min)		
máximo (max)		

que la correlación entre grupos. Por ejemplo, todas las correlaciones dentro del grupo, para el Grupo 1, son mayores que 0.9 (excepto para $\text{corr}(min, q3) = 0.89$). Las características del Grupo 3 son las menos correlacionadas.

Una vez determinados los grupos, la selección de características se realizó mediante variables de diferentes grupos. De esta forma evitamos la multicolinealidad, que puede subestimar la importancia de algunas características. Además, limitar a 3 variables (una de cada grupo) puede reducir el costo computacional y energía requerida por los dispositivos. Evaluamos combinaciones de 1, 2 y 3 características de la siguiente manera: cuando se usa una característica única es posible elegir cualquier característica de cualquier grupo. Cuando se utilizan 2 caracteristicas, es posible elegir 1 característica de posición y 1 característica de dispersión o 1 de posición y 1 de forma o 1 de dispersión y 1 de forma. Cuando se usan 3 características se selecciona una característica de cada grupo.

Teniendo en cuenta este criterio, todas las combinaciones de 1, 2 y 3 características fueron evaluados para los 4 modelos diferentes: Discriminante Lineal (LDA), Discriminante Cuadrático (QDA), Regresión Logística (LR), y K Vecinos mas próximos (KNN). Se consideró la precisión como medida de comparación. Recordemos que la precisión se define como el porcentaje de correctamente clasificados: (TP + TN) / (TP + TN + FP + FN), donde TP, FN, FP y TN representan el número de verdaderos positivos, falsos negativos, falsos positivos y verdaderos negativos, respectivamente.

Para obtener estimaciones correctas de la precisión de los modelos, se entrenó y evaluó cada modelo utilizando validación cruzada con $KFold = 10$ y con 5 repeticiones. Por lo tanto, cada puntuación de precisión calculada en el conjunto de validación se establece como el promedio de 50 valores.

8.2 Resultados

La estimación de proximidad se analizó en 2 escenarios diferentes. Cada uno de ellos consideran que las aplicaciones de rastreo de contactos pueden ser conscientes de su entorno para determinar si un contacto es cercano o no. En estos escenarios, las características se seleccionan para optimizar la precisión de un conjunto de datos específico relacionado con un ambiente. Por lo tanto, las características seleccionadas pueden ser diferentes para cada entorno, lo que significa que estas aplicaciones pueden tener dos modelos diferentes: uno para estimación de interiores, y otra para exterior. Si bien, la aplicación no conoce en qué entorno se encuentran los dispositivos, entrenar modelos diferenciales es una primera aproximación al problema. Esto es, una vez calculada la precisión de estos modelos se podría comparar con la precisión de un modelo que no diferencie entre indoor y outdoor y así evaluar el aporte de incorporar un mecanismo que permita a la app distinguir entre ambientes. No abarcaremos este problema en este trabajo.

Para cada entorno, también evaluamos la ganancia de introducir múltiples características en los modelos. Para ello, evaluamos el desempeño ganancia de agregar hasta tres características, una de cada grupo como se discutió anteriormente. Incluso si esto permite comparar la ganancia relativa de introducir una nueva característica, no proporciona una comparación absoluta para evaluar su rendimiento real. Las distancias de hasta 2 metros se consideran contactos cercanos para calcular la precisión general y comparar con resultados de los modelos de clasificación.

8.2.1 Análisis en interiores:

Primero, analizamos la estimación de proximidad en ambientes interiores. Consideramos diferentes escenarios definidos por los modelos de clasificación y por el número de características consideradas (de 1 a 3). Para cada escenario, los 5 mejores conjuntos de características, en términos de precisión, fueron seleccionados para mostrar.

La Figura 8.2 resume el rendimiento para los 12 escenarios diferentes resultantes de los 4 tipos de modelos y los 3 recuentos de características. Se muestra la precisión para la selección de cada característica utilizando diagramas de caja ordenados por su valor medio de izquierda a derecha. Para todos los casos, los resultados confirman que la precisión no cambia considerablemente a medida que usamos más características en los modelos. Con respecto a la variabilidad de la precisión, es casi idéntica para cualquiera de las cantidades de características consideradas. Se observa que la variable max es la que más se repite entre todos los escenarios. Además, notamos que las variables del Grupo 1 aparecen en todas las combinaciones de 2 variables, lo que significa que podemos evitar el caso de 1 variable de dispersión y 1 variable de forma.

8.2.2 Análisis al aire libre

A continuación, consideramos la estimación de proximidad en entornos al aire libre. Similar a los escenarios para el caso interior se muestran en la Figura 8.3. Si se compara con las Figura 8.2, se hace evidente que la estimación de proximidad en exteriores puede ser mucho menos precisa que para espacios interiores. Sin embargo, todos tienen un rendimiento bastante similar y sigue presente la misma tendencia que en nuestro análisis anterior: la precisión no cambia considerablemente a medida que usamos más características en los modelos y la variabilidad de la precisión es casi identica para cualquiera de las cantidades de características consideradas. Al igual que en ambientes interiores, la variable max aparece mas veces en todos los modelos de todos los escenarios considerados.

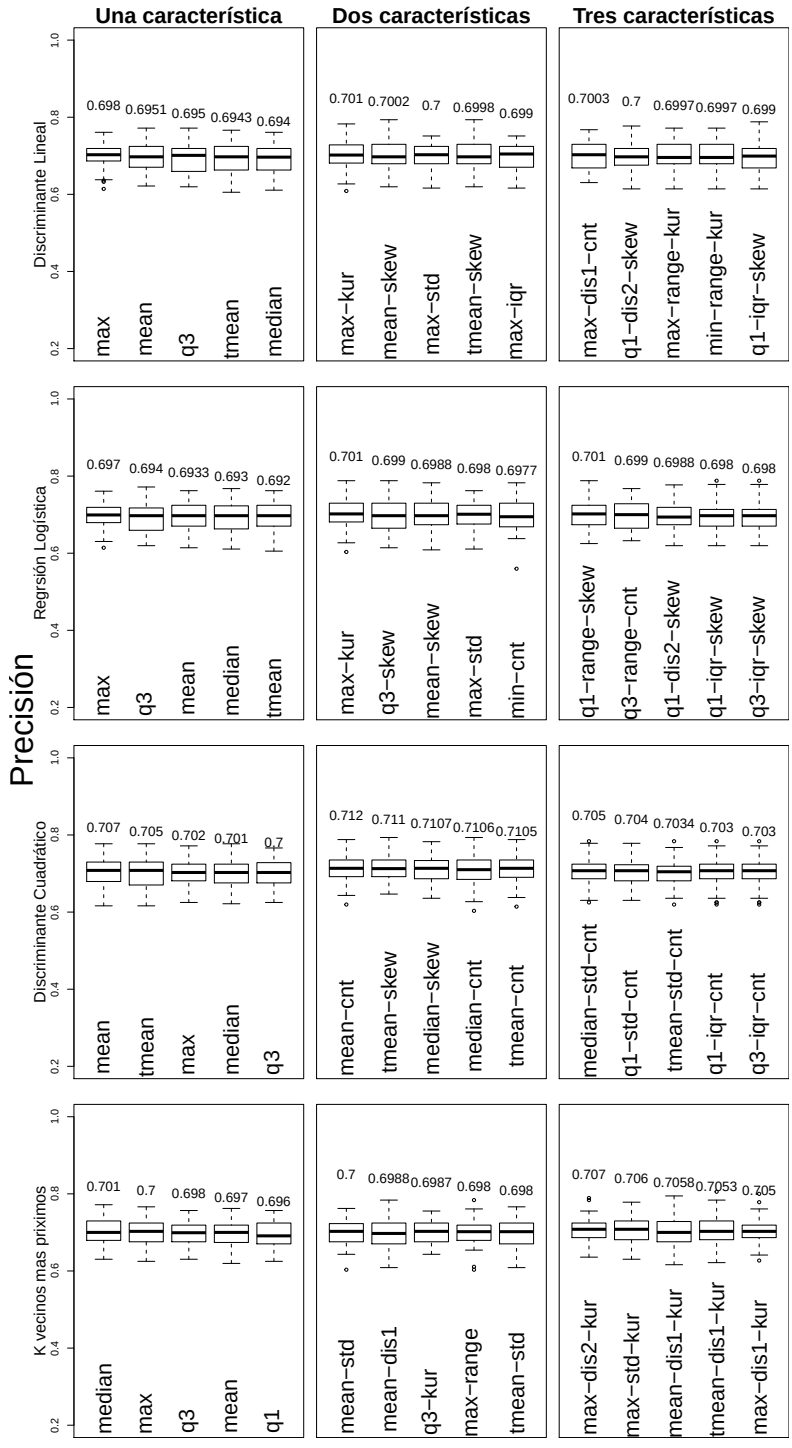


Figura 8.2: Precisión en el entorno interior con respecto al número de características consideradas para diferentes modelos. Primera fila: Modelos de LDA. Segunda fila: Modelos de regresión logística. Tercera fila: Modelos de QDA. Cuarta fila: Modelos de KNN.

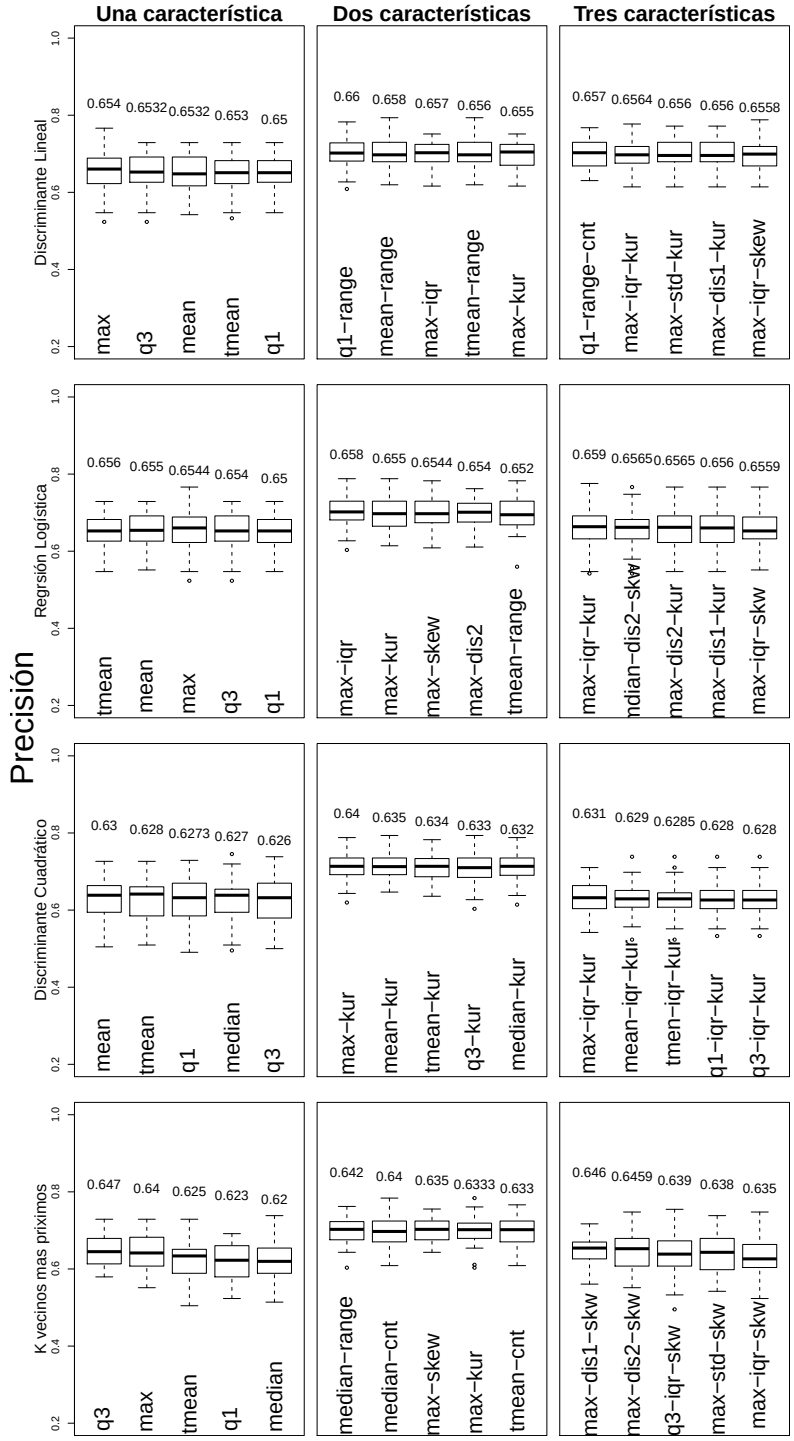


Figura 8.3: Precisión en el entorno exterior con respecto al número de características consideradas para diferentes modelos. Primera fila: Modelos de LDA. Segunda fila: Modelos de regresión logística. Tercera fila: Modelos de QDA. Cuarta fila: Modelos de KNN.

8.3 Conclusiones

Además de comparar el rendimiento de cada método de clasificación, nos centramos en el proceso de selección de características con el objetivo de comprender qué datos se pueden utilizar para mejorar la precisión.

Los resultados obtenidos demostraron que aumentar el recuento de características no contribuye a mejorar la estimación de proximidad para ninguno de los dos ambientes. Además, consideramos que al no identificar diferencias en la precisión de cada método, utilizar cualquiera de estos métodos no cambiará sustancialmente la estimación de proximidad de la aplicación. Cabe aclarar que pueden existir otros métodos, no estudiados en este trabajo, que sí proporcionen evidencias de mejorar la precisión al estimar la proximidad. Por ejemplo, podríamos probar métodos de clasificación cómo RF (Random Forests) James et al. (2013) (página 303) o SVM (support vector machines) James et al. (2013) (página 337), pero estos análisis se escapan a los objetivos de este trabajo.

Apéndice A

8.4 Pruebas de normalidad.

8.4.1 Métodos gráficos.

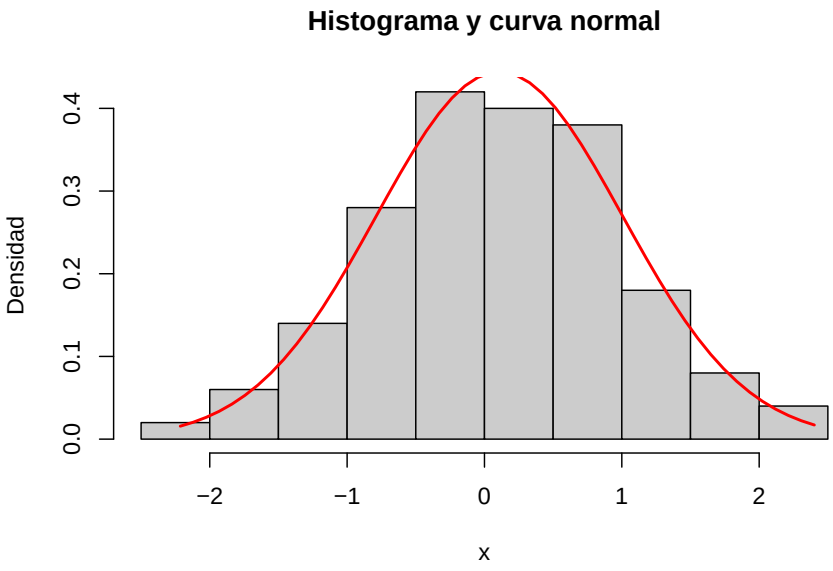


Figura 8.4: Ejemplo de cómo se tiene que ver un histograma de una variable que sigue una distribución normal

En la figura 8.4 vemos un histograma de una muestra que fue extraída de una variable normal y superponemos la curva de la densidad teórica de donde sacamos la muestra. Si las barras varían considerablemente respecto de la curva, los datos podrían no ser normales.

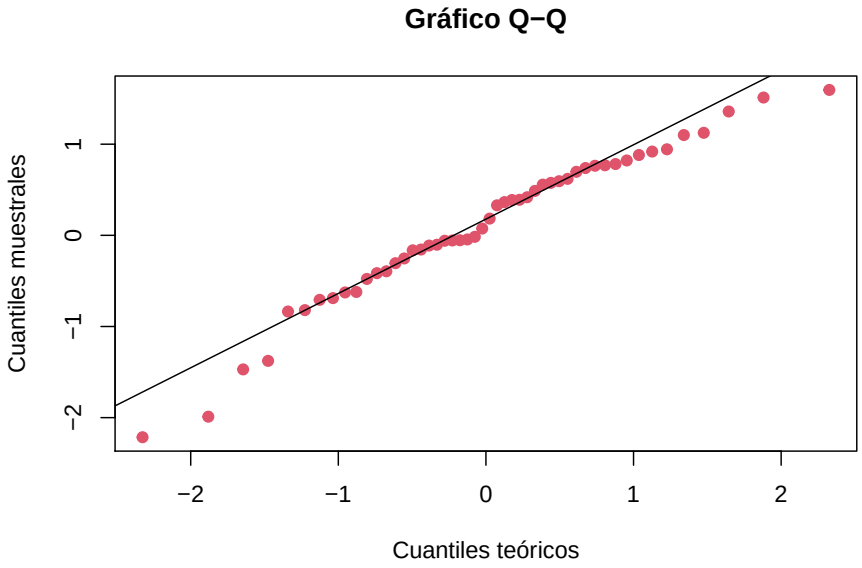


Figura 8.5: Ejemplo de un gráfico Q-Q normal de datos N(0,1) generados aleatoriamente.

En la figura 8.5 vemos un gráfico de cuantiles la misma muestra que utilizamos en la figura 8.4. Si las puntos varían considerablemente respecto de la recta, los datos podrían no ser normales.

8.4.2 Valores atípicos multivariantes (Multivariate outliers)

Los valores atípicos multivariados son la razón común para violar la suposición de normalidad multivariada. En otras palabras, la suposición de normalidad multivariada requiere la ausencia de valores atípicos multivariados. Por lo tanto, es crucial verificar si los datos tienen valores atípicos multivariantes, antes de comenzar con el análisis multivariado. El paquete MVN incluye dos métodos de detección de valores atípicos multivariados que se basan en distancias robustas de Mahalanobis (rMD (x)). La distancia de Mahalanobis es una métrica que calcula qué tan lejos está cada observación del centro de distribución de la articulación, que se puede considerar como el centroide en el espacio multivariable. Las distancias robustas se estiman a partir de estimadores determinantes de covarianza mínima en lugar de la covarianza de la muestra. Estos dos enfoques, definidos como la distancia Mahalanobis y la distancia Mahalanobis ajustada en el paquete, detectan valores atípicos multivariados como se indica a continuación,

Distancia Mahalanobis:

1 Calcule distancias robustas de Mahalanobis (rMD (xi)),

2 Calcule el cuantil del 97.5 por ciento (Q) de la distribución de chi-cuadrado,

3 Declare rMD (xi)> Q como posible valor atípico.

Distancia ajustada de Mahalanobis:

1 Calcule distancias robustas de Mahalanobis (rMD (xi)),

2 Calcule el cuantil ajustado de 97.5 por ciento (AQ) de la distribución de chi-Cuadrado,

3 Declare rMD (xi)> AQ como posible valor atípico.

8.5 Shapiro Wilk

No existe un contraste “óptimo” para probar la hipótesis de normalidad. La razón es que la potencia relativa depende del tamaño muestral y de la verdadera distribución que genera los datos. Desde un punto de vista poco riguroso, el contraste de Shapiro y Wilks es, en términos generales, el más conveniente en pequeñas muestras ($n < 30$), este contraste mide el ajuste de la muestra representada en papel probabilístico normal a una recta. Se rechaza la

normalidad cuando el ajuste es malo, que corresponde a valores pequeños del estadístico

8.6 Mardia, Henze-Zirkler y Royston

La prueba de Henze-Zirkler está basada en la distancia funcional no negativa, la cual mide la distancia entre dos funciones de distribución. Si los datos presentan una distribución normal multivariada, la prueba estadística se distribuye aproximadamente como una lognormal. Primero, la media, varianza y el parámetro de suavización son calculados. Entonces, la media y la varianza son lognormalizados y el pvalor es estimado.

La prueba de Royston usa la estadística Shapiro-Wilk / Shapiro-Francia para probar la normalidad multivariada. Si la curtosis es mayor a 3, entonces se usa la prueba de Shapiro-Francia para distribuciones leptocurticas. Mientras que se usa la prueba de Shapiro-Wilk para distribuciones platicurticas.

Para mas detalles sobre estos métodos, se puede buscar en Henze and Zirkler (1990) y Royston (1983).

8.7 Test de Barlett y Box M

De entre las diferentes pruebas que contrastan la homogeneidad de varianza, el más recomendable cuando solo hay un predictor, dado que se asume que se distribuye de forma normal, es el test de Barttlet que fué desarrollado por el matemático Bartlett (1937). Cuando se emplean múltiples predictores, se tiene que contrastar que la matriz de covarianzas (Σ) es constante en todos los grupos, siendo recomendable comprobar también la homogeneidad de varianza para cada predictor a nivel individual.

El test Box M fué desarrollado por el matemático Box (1949) como una extensión del test de Barttlet para escenarios multivariantes y permite contrastar la igualdad de matrices entre grupos. El test Box M es muy sensible a violaciones de la normalidad multivariante, por lo que ésta debe ser contrastada con anterioridad. Ocurre con frecuencia, que el resultado de un test Box M resulta significativo debido a la falta de distribución normal multivariante en lugar de por falta de homogeneidad en las matrices de covarianza. Dada la sensibilidad de este test se recomienda emplear un límite de significancia de 0.001

Apéndice B

El siguiente código contiene la función que utilizamos para elegir K usando validación cruzada mencionada en el capítulo 7.

```
KNN.cv <- function(Train, Test, Cl) {

  # Dividir el conjunto de datos en conjuntos de entrenamiento
  # y prueba de forma aleatoria, pero necesitamos establecer
  # la semilla para generar el mismo valor cada vez que ejecutamos
  # el código

  set.seed(1)

  # Crear un índice para dividir los datos: 80% de entrenamiento
  # y 20% de prueba

  index = round(nrow(Train)*0.2,digits=0)

  # Muestrear aleatoriamente en todo el conjunto de datos y
  # mantener el número total igual al valor del índice

  test.indices = sample(1:nrow(Train), index)

  # 80% training set

  train=Train[-test.indices,]
  train.clases=Cl[-test.indices]

  # 20% test set

  test=Train[test.indices,]
  test.clases=Cl[test.indices]

  if (round(sqrt(dim(Train)[1]),0) < 10) {
    K <- 1:9
```

```
    }else{K <- c(1:9, seq(10, round(sqrt(dim(Train)[1]),0), 10))}

  error     <- rep(0, length(K))
  error.index=1
  for (i in K) {

    knn.class <- knn(data.frame(train), data.frame(test),
                     train.clases, k = i)
    error[error.index] <- mean(knn.class != test.clases)
    error.index=error.index+1
  }

  j <- K[which.min(error)]
  return(knn(Train, Test, Cl, k = j))
}
```

Este código es el que utilizamos en los 7 escenarios de capítulo 7, el lugar donde creamos los datos de prueba y de entrenamiento lo dejamos en blanco para ser rellenados dependiendo del escenario. Al final de todo se encuentran los códigos dónde creamos los datos.

```
# CARGAMOS LIBRERIA NECESARIA
#-----------------------------------------------------------------------

library (MASS)   ## En esta librería están las funciones LDA y QDA.
library (class)  ## En esta librería están las funciones de K-NN.
library(mvtnorm) ## En esta librería están las funciones para
                 ## crear datos proveniente de una distribución
                 ## normal multivariada.

# CREAMOS VECTORES QUE MOSTRAREMOS AL FINAL DE LA SIMULACION
#-----------------------------------------------------------------------

TP.lda   <- rep(0, 100)
TN.lda   <- rep(0, 100)
FP.lda   <- rep(0, 100)
FN.lda   <- rep(0, 100)
ACC.lda  <- rep(0, 100)
ERROR.lda<- rep(0, 100)
SENS.lda <- rep(0, 100)
ESP.lda  <- rep(0, 100)
PREC.lda <- rep(0, 100)

TP.qda   <- rep(0, 100)
TN.qda   <- rep(0, 100)
```

```
FP.qda   <- rep(0, 100)
FN.qda   <- rep(0, 100)
ACC.qda  <- rep(0, 100)
ERROR.qda<- rep(0, 100)
SENS.qda <- rep(0, 100)
ESP.qda  <- rep(0, 100)
PREC.qda <- rep(0, 100)

TP.glm   <- rep(0, 100)
TN.glm   <- rep(0, 100)
FP.glm   <- rep(0, 100)
FN.glm   <- rep(0, 100)
ACC.glm  <- rep(0, 100)
ERROR.glm<- rep(0, 100)
SENS.glm <- rep(0, 100)
ESP.glm  <- rep(0, 100)
PREC.glm <- rep(0, 100)

TP.knn1   <- rep(0, 100)
TN.knn1   <- rep(0, 100)
FP.knn1   <- rep(0, 100)
FN.knn1   <- rep(0, 100)
ACC.knn1  <- rep(0, 100)
ERROR.knn1<- rep(0, 100)
SENS.knn1 <- rep(0, 100)
ESP.knn1  <- rep(0, 100)
PREC.knn1 <- rep(0, 100)

TP.knnCV   <- rep(0, 100)
TN.knnCV   <- rep(0, 100)
FP.knnCV   <- rep(0, 100)
FN.knnCV   <- rep(0, 100)
ACC.knnCV  <- rep(0, 100)
ERROR.knnCV<- rep(0, 100)
SENS.knnCV <- rep(0, 100)
ESP.knnCV  <- rep(0, 100)
PREC.knnCV <- rep(0, 100)


set.seed(100)


# CREO DATOS DE TEST
#-------------------------------------------------------------
```

```
## Ver escenario X.

#### ------------ COMIENZA EL CICLO ------------------- ###
for (i in 1:100) {


# CREO DATOS DE ENTRENAMIENTO
#---------------------------------------------------------------------

  ## Ver escenario X.

# MODELO
#---------------------------------------------------------------------

lda.mod <- lda(clases ~. , data = Datos)
qda.mod <- qda(clases ~. , data = Datos)
glm.mod <- glm(clases ~. , data = Datos, family = "binomial")


# TESTEO CON LOS DATOS DE PRUEBA
#---------------------------------------------------------------------

lda.pred  <- predict(lda.mod, Datos.test)
qda.pred  <- predict(qda.mod, Datos.test)
glm.probs <- predict(glm.mod, Datos.test, type="response")

# CLASES Y TP, TN, FP, FN, ACC, PREC
#---------------------------------------------------------------------

lda.class <- lda.pred$class
qda.class <- qda.pred$class
glm.class <- rep(0 ,dim(Datos.test)[1])
glm.class[glm.probs >.5] <- 1
knn1.class  <- knn(Datos[,1:2], Datos.test[,1:2], Datos$clases,
                   k = 1)
knnCV.class <- KNN.cv(Datos[,1:2], Datos.test[,1:2], Datos$clases,
                      Datos.test$clases)

lda.tabla <- table(lda.class , Datos.test$clases)

TP.lda[i]   <- lda.tabla[1,1]
TN.lda[i]   <- lda.tabla[2,2]
FP.lda[i]   <- lda.tabla[2,1]
FN.lda[i]   <- lda.tabla[1,2]
ACC.lda[i]  <- (TP.lda[i] + TN.lda[i]) / (TP.lda[i] + TN.lda[i]
```

```
                                          + FP.lda[i] + FN.lda[i])
ERROR.lda[i]<- (FP.lda[i] + FN.lda[i]) / (TP.lda[i] + TN.lda[i]
                                          + FP.lda[i] + FN.lda[i])
PREC.lda[i] <- TP.lda[i] / (TP.lda[i] + FP.lda[i])
SENS.lda[i] <- TP.lda[i] / (TP.lda[i] + FN.lda[i])
ESP.lda[i]  <- TN.lda[i] / (TN.lda[i] + FP.lda[i])

qda.tabla <- table(qda.class , Datos.test$clases)

TP.qda[i]   <- qda.tabla[1,1]
TN.qda[i]   <- qda.tabla[2,2]
FP.qda[i]   <- qda.tabla[2,1]
FN.qda[i]   <- qda.tabla[1,2]
ACC.qda[i]  <- (TP.qda[i] + TN.qda[i]) / (TP.qda[i] + TN.qda[i] +
                                            FP.qda[i] + FN.qda[i])
ERROR.qda[i]<- (FP.qda[i] + FN.qda[i]) / (TP.qda[i] + TN.qda[i] +
                                            FP.qda[i] + FN.qda[i])
PREC.qda[i] <- TP.qda[i] / (TP.qda[i] + FP.qda[i])
SENS.qda[i] <- TP.qda[i] / (TP.qda[i] + FN.qda[i])
ESP.qda[i]  <- TN.qda[i] / (TN.qda[i] + FP.qda[i])

glm.tabla <- table(glm.class , Datos.test$clases)

TP.glm[i]   <- glm.tabla[1,1]
TN.glm[i]   <- glm.tabla[2,2]
FP.glm[i]   <- glm.tabla[2,1]
FN.glm[i]   <- glm.tabla[1,2]
ACC.glm[i]  <- (TP.glm[i] + TN.glm[i]) / (TP.glm[i] + TN.glm[i] +
                                            FP.glm[i] + FN.glm[i])
ERROR.glm[i]<- (FP.glm[i] + FN.glm[i]) / (TP.glm[i] + TN.glm[i] +
                                            FP.glm[i] + FN.glm[i])
PREC.glm[i] <- TP.glm[i] / (TP.glm[i] + FP.glm[i])
SENS.glm[i] <- TP.glm[i] / (TP.glm[i] + FN.glm[i])
ESP.glm[i]  <- TN.glm[i] / (TN.glm[i] + FP.glm[i])

knn1.tabla <- table(knn1.class , Datos.test$clases)

TP.knn1[i]   <- knn1.tabla[1,1]
TN.knn1[i]   <- knn1.tabla[2,2]
FP.knn1[i]   <- knn1.tabla[2,1]
FN.knn1[i]   <- knn1.tabla[1,2]
ACC.knn1[i]  <- (TP.knn1[i] + TN.knn1[i]) / (TP.knn1[i] + TN.knn1[i] +
                                               FP.knn1[i] + FN.knn1[i])
ERROR.knn1[i]<- (FP.knn1[i] + FN.knn1[i]) / (TP.knn1[i] + TN.knn1[i] +
                                               FP.knn1[i] + FN.knn1[i])
```

```
PREC.knn1[i] <- TP.knn1[i] / (TP.knn1[i] + FP.knn1[i])
SENS.knn1[i] <- TP.knn1[i] / (TP.knn1[i] + FN.knn1[i])
ESP.knn1[i]  <- TN.knn1[i] / (TN.knn1[i] + FP.knn1[i])

knnCV.tabla <- table(knnCV.class , Datos.test$clases)

TP.knnCV[i]   <- knnCV.tabla[1,1]
TN.knnCV[i]   <- knnCV.tabla[2,2]
FP.knnCV[i]   <- knnCV.tabla[2,1]
FN.knnCV[i]   <- knnCV.tabla[1,2]
ACC.knnCV[i]  <- (TP.knnCV[i] + TN.knnCV[i]) / (TP.knnCV[i] + TN.knnCV[i]
                                                + FP.knnCV[i] + FN.knnCV[i])
ERROR.knnCV[i]<- (FP.knnCV[i] + FN.knnCV[i]) / (TP.knnCV[i] + TN.knnCV[i]
                                                + FP.knnCV[i] + FN.knnCV[i])
PREC.knnCV[i] <- TP.knnCV[i] / (TP.knnCV[i] + FP.knnCV[i])
SENS.knnCV[i] <- TP.knnCV[i] / (TP.knnCV[i] + FN.knnCV[i])
ESP.knnCV[i]  <- TN.knnCV[i] / (TN.knnCV[i] + FP.knnCV[i])

}

#### ------------ TERMINA EL CICLO ------------------- ###
```

```
## Escenario 1
# CREO DATOS DE TEST
#-----------------------------------------------------------------

datos1.test <- rmvnorm(200, mean = c(0, 0),
                       sigma = matrix(c(1, 0, 0, 1), nrow = 2))
datos1.test <- cbind(datos1.test, rep(0,dim(datos1.test)[1]))
datos2.test <- rmvnorm(200, mean = c(1, 1),
                       sigma = matrix(c(1, 0, 0, 1), nrow = 2))
datos2.test <- cbind(datos2.test, rep(1,dim(datos2.test)[1]))

Datos.test <- rbind(datos1.test, datos2.test)
Datos.test <- Datos.test[sample(1:dim(Datos.test)[1]),]
Datos.test <- as.data.frame(Datos.test)
colnames(Datos.test) <- c("x1", "x2", "clases")

# CREO DATOS DE ENTRENAMIENTO
#-----------------------------------------------------------------

datos1 <- rmvnorm(20, mean = c(0, 0),
                  sigma = matrix(c(1, 0, 0, 1), nrow = 2))
datos1 <- cbind(datos1, rep(0,dim(datos1)[1]))
datos2 <- rmvnorm(20, mean = c(1, 1),
```

```
                  sigma = matrix(c(1, 0, 0, 1), nrow = 2))
datos2 <- cbind(datos2, rep(1,dim(datos2)[1]))

Datos <- rbind(datos1, datos2)
Datos <- Datos[sample(1:dim(Datos)[1]),]
Datos <- as.data.frame(Datos)
colnames(Datos) <- c("x1", "x2", "clases")
```

```
## Escenario 2
# CREO DATOS DE TEST
#-----------------------------------------------------------------

datos1.test <- rmvnorm(200, mean = c(0, 0),
                       sigma = matrix(c(1, 0.5, 0.5, 1), nrow = 2))
datos1.test <- cbind(datos1.test, rep(0,dim(datos1.test)[1]))
datos2.test <- rmvnorm(200, mean = c(1, 1),
                       sigma = matrix(c(1, 0.5, 0.5, 1), nrow = 2))
datos2.test <- cbind(datos2.test, rep(1,dim(datos2.test)[1]))

Datos.test <- rbind(datos1.test, datos2.test)
Datos.test <- Datos.test[sample(1:dim(Datos.test)[1]),]
Datos.test <- as.data.frame(Datos.test)
colnames(Datos.test) <- c("x1", "x2", "clases")

# CREO DATOS DE ENTRENAMIENTO
#-----------------------------------------------------------------

datos1 <- rmvnorm(20, mean = c(0, 0),
                  sigma = matrix(c(1, 0.5, 0.5, 1), nrow = 2))
datos1 <- cbind(datos1, rep(0,dim(datos1)[1]))
datos2 <- rmvnorm(20, mean = c(1, 1),
                  sigma = matrix(c(1, 0.5, 0.5, 1), nrow = 2))
datos2 <- cbind(datos2, rep(1,dim(datos2)[1]))

Datos <- rbind(datos1, datos2)
Datos <- Datos[sample(1:dim(Datos)[1]),]
Datos <- as.data.frame(Datos)
colnames(Datos) <- c("x1", "x2", "clases")
```

```
## Escenario 3
# CREO DATOS DE TEST
#-----------------------------------------------------------------

datos1.test <- rmvt(200, df = 5, delta = c(0, 0))
datos1.test <- cbind(datos1.test, rep(0,dim(datos1.test)[1]))
```

```
datos2.test <- rmvt(200, df = 5, delta = c(0, 1))
datos2.test <- cbind(datos2.test, rep(1,dim(datos2.test)[1]))

Datos.test <- rbind(datos1.test, datos2.test)
Datos.test <- Datos.test[sample(1:dim(Datos.test)[1]),]
Datos.test <- as.data.frame(Datos.test)
colnames(Datos.test) <- c("x1", "x2", "clases")

# CREO DATOS DE ENTRENAMIENTO
#----------------------------------------------------------------

datos1 <- rmvt(20, df = 5, delta = c(0, 0))
datos1 <- cbind(datos1, rep(0,dim(datos1)[1]))
datos2 <- rmvt(20, df = 5, delta = c(0, 1))
datos2 <- cbind(datos2, rep(1,dim(datos2)[1]))

Datos <- rbind(datos1, datos2)
Datos <- Datos[sample(1:dim(Datos)[1]),]
Datos <- as.data.frame(Datos)
colnames(Datos) <- c("x1", "x2", "clases")
```

```
## Escenario 4
# CREO DATOS DE TEST
#----------------------------------------------------------------

  datos1.test <- matrix(rchisq(100, df = 3, ncp = 1), ncol = 2)
  datos1.test <- cbind(datos1.test, rep(0,dim(datos1.test)[1]))
  datos2.test <- matrix(rchisq(100, df = 3, ncp = 0), ncol = 2)
  datos2.test <- cbind(datos2.test, rep(1,dim(datos2.test)[1]))

  Datos.test <- rbind(datos1.test, datos2.test)
  Datos.test <- Datos.test[sample(1:dim(Datos.test)[1]),]
  Datos.test <- as.data.frame(Datos.test)
  colnames(Datos.test) <- c("x1", "x2", "clases")

# CREO DATOS DE ENTRENAMIENTO
#----------------------------------------------------------------

  datos1 <- matrix(rchisq(20, df = 3, ncp = 1), ncol = 2)
  datos1 <- cbind(datos1, rep(0,dim(datos1)[1]))
  datos2 <- matrix(rchisq(20, df = 3, ncp = 0), ncol = 2)
  datos2 <- cbind(datos2, rep(1,dim(datos2)[1]))

  Datos <- rbind(datos1, datos2)
  Datos <- Datos[sample(1:dim(Datos)[1]),]
```

```
  Datos <- as.data.frame(Datos)
  colnames(Datos) <- c("x1", "x2", "clases")
```

```
## Escenario 5
# CREO DATOS DE TEST
#-----------------------------------------------------------------

datos1.test <- rmvnorm(100, mean = c(0, 0),
                       sigma = matrix(c(1, 0, 0, 1), nrow = 2))
datos1.test[1:dim(datos1.test)[1]/2, ] <-
  datos1.test[1:dim(datos1.test)[1]/2, ] + transl
datos1.test[(dim(datos1.test)[1]/2 + 1):dim(datos1.test)[1], ] <- datos1.test[(dim(datos1.test)[1
datos1.test <- cbind(datos1.test, rep(0,dim(datos1.test)[1]))
datos2.test <- rmvnorm(100, mean = c(0, 0),
                       sigma = matrix(c(1, 0, 0, 1), nrow = 2))
datos2.test <- cbind(datos2.test, rep(1,dim(datos2.test)[1]))

Datos.test <- rbind(datos1.test, datos2.test)
Datos.test <- Datos.test[sample(1:dim(Datos.test)[1]),]
Datos.test <- as.data.frame(Datos.test)
colnames(Datos.test) <- c("x1", "x2", "clases")

# CREO DATOS DE ENTRENAMIENTO
#-----------------------------------------------------------------

datos1 <- rmvnorm(40, mean = c(0, 0),
                  sigma = matrix(c(1, 0, 0, 1), nrow = 2))
datos1[1:dim(datos1)[1]/2, ] <-
  datos1[1:dim(datos1)[1]/2, ] + transl
datos1[(dim(datos1)[1]/2 + 1):dim(datos1)[1], ] <-
  datos1[(dim(datos1)[1]/2 + 1):dim(datos1)[1], ] - transl
datos1 <- cbind(datos1, rep(0,dim(datos1)[1]))
datos2 <- rmvnorm(40, mean = c(0, 0),
                  sigma = matrix(c(1, 0, 0, 1), nrow = 2))
datos2 <- cbind(datos2, rep(1,dim(datos2)[1]))

Datos <- rbind(datos1, datos2)
Datos <- Datos[sample(1:dim(Datos)[1]),]
Datos <- as.data.frame(Datos)
colnames(Datos) <- c("x1", "x2", "clases")
```

```
## Escenario 6
# CREO DATOS DE TEST
#-----------------------------------------------------------------
```

```
datos1.test <- rmvnorm(200, mean = c(0, 0),
                       sigma = matrix(c(1, 0.5, 0.5, 1), nrow = 2))
datos1.test <- cbind(datos1.test, rep(0,dim(datos1.test)[1]))
datos2.test <- rmvnorm(200, mean = c(1, 1),
                       sigma = matrix(c(1, -0.5, -0.5, 1), nrow = 2))
datos2.test <- cbind(datos2.test, rep(1,dim(datos2.test)[1]))

Datos.test <- rbind(datos1.test, datos2.test)
Datos.test <- Datos.test[sample(1:dim(Datos.test)[1]),]
Datos.test <- as.data.frame(Datos.test)
colnames(Datos.test) <- c("x1", "x2", "clases")

# CREO DATOS DE ENTRENAMIENTO
#-----------------------------------------------------------------

datos1 <- rmvnorm(20, mean = c(0, 0),
                  sigma = matrix(c(1, 0.5, 0.5, 1), nrow = 2))
datos1 <- cbind(datos1, rep(0,dim(datos1)[1]))
datos2 <- rmvnorm(20, mean = c(1, 1),
                  sigma = matrix(c(1, -0.5, -0.5, 1), nrow = 2))
datos2 <- cbind(datos2, rep(1,dim(datos2)[1]))

Datos <- rbind(datos1, datos2)
Datos <- Datos[sample(1:dim(Datos)[1]),]
Datos <- as.data.frame(Datos)
colnames(Datos) <- c("x1", "x2", "clases")
```

```
## Escenario 7
# CREO DATOS DE TEST
#-----------------------------------------------------------------
  r1 <- runif(50)
  t1 <- 2*pi*runif(50)
  r2 <- runif(100, 1.8,2.2)
  t2 <- 2*pi*runif(100)
  r3 <- runif(50, 2.9,3.4)
  t3 <- 2*pi*runif(50)

  x1 <- r1*cos(t1)+rnorm(50,0, 0.15)
  y1 <- r1*sin(t1)+rnorm(50,0, 0.15)

  x2 <- r2*cos(t2)+rnorm(100,0, 0.15)
  y2 <- r2*sin(t2)+rnorm(100,0, 0.15)

  x3 <- r3*cos(t3)+rnorm(50,0, 0.15)
  y3 <- r3*sin(t3)+rnorm(50,0, 0.15)
```

```
  datos1.test <- cbind(c(x1,x3), c(y1, y3))
  datos1.test <- cbind(datos1.test, rep(0,dim(datos1.test)[1]))
  datos2.test <- cbind(x2,y2)
  datos2.test <- cbind(datos2.test, rep(1,dim(datos2.test)[1]))

  Datos.test <- rbind(datos1.test, datos2.test)
  Datos.test <- Datos.test[sample(1:dim(Datos.test)[1]),]
  Datos.test <- as.data.frame(Datos.test)
  colnames(Datos.test) <- c("x1", "x2", "clases")

  # CREO DATOS DE ENTRENAMIENTO
#------------------------------------------------------------------

  r1 <- runif(20)
  t1 <- 2*pi*runif(20)
  r2 <- runif(40, 1.8,2.2)
  t2 <- 2*pi*runif(40)
  r3 <- runif(20, 2.9,3.4)
  t3 <- 2*pi*runif(20)

  x1 <- r1*cos(t1)+rnorm(20,0, 0.15)
  y1 <- r1*sin(t1)+rnorm(20,0, 0.15)

  x2 <- r2*cos(t2)+rnorm(40,0, 0.15)
  y2 <- r2*sin(t2)+rnorm(40,0, 0.15)

  x3 <- r3*cos(t3)+rnorm(20,0, 0.15)
  y3 <- r3*sin(t3)+rnorm(20,0, 0.15)

  datos1 <- cbind(c(x1,x3), c(y1, y3))
  datos1 <- cbind(datos1, rep(0,dim(datos1)[1]))
  datos2 <- cbind(x2,y2)
  datos2 <- cbind(datos2, rep(1,dim(datos2)[1]))

  Datos <- rbind(datos1, datos2)
  Datos <- Datos[sample(1:dim(Datos)[1]),]
  Datos <- as.data.frame(Datos)
  colnames(Datos) <- c("x1", "x2", "clases")
```

Bibliografía

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 160(901):268–282.

Box, G. E. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika*, 36(3/4):317–346.

da Silva, A. R. (2017). Tools for biometry and applied statistics in agricultural science. *The comprehensive R archive network. Website https://CRAN. R-project. org/package= psych*.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70(352):892–898.

Fox, J. (1984). *Linear statistical models and related methods: With applications to social research*. Number 1. New York; Toronto: Wiley.

Hand, D. J. and Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3):523–541.

Henze, N. and Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Communications in statistics-Theory and Methods*, 19(10):3595–3617.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Korkmaz, S., Goksuluk, D., and Zararsiz, G. (2014). Mvn: An r package for assessing multivariate normality. *The R Journal*, 6(2):151–162.

Maddala, G. (1983). Methods of estimation for models of markets with bounded price variation. *International Economic Review*, pages 361–378.

Peña, D. (2001). Fundamentos de estadística, alianza editorial.

Peña, D. (2002). *Análisis de datos multivariantes*, volume 24. McGraw-hill Madrid.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rappaport, T. S. et al. (1996). *Wireless communications: principles and practice*, volume 2. prentice hall PTR New Jersey.

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., Firth, D., and Ripley, M. B. (2013). Package 'mass'. *Cran r*, 538:113–120.

Rosenberg, E. and Gleit, A. (1994). Quantitative methods in credit management: a survey. *Operations research*, 42(4):589–613.

Royston, J. (1983). Some techniques for assessing multivarate normality based on the shapiro-wilk w. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 32(2):121–133.