UNIVERSIDAD NACIONAL DE COLOMBIA

PREGRADO EN ESTADISTICA

Departamento de Estadística
Facultad de Ciencias

— Introducción al Análisis Multivariado —

Analisis de Discriminantes
Multivariados
Datos: Datos pacientes Anemia
tarea número 7

Integrantes:

Ricardo William Salazar Espinal C.C. 1017219472

Medellín, Colombia

Medellin, septiemnre 13 de 2024

Índice

Índice de figuras

Índice de cuadros

1 Introduccion

El análisis discriminante tiene como propósito desarrollar una función o regla que permita clasificar a individuos u observaciones en diferentes grupos o categorías. Basándose en los valores multivariados de ciertas variables $(X_1, X_2, \ldots, X_p)$, la técnica busca determinar a qué grupo o categoría pertenece un individuo dado, representado por el vector $\underline{\mathbf{x}}$. En esencia, el análisis discriminante utiliza estos puntajes para asignar a cada observación a su grupo más probable.

Para este analisis consideraremos una base de datos de personas que se realizaron estudios para determinar si padecen de anemia, la base de datos cuenta con un total de 104 observaciones y 7 variables las cuales son:

Number: variable que dice que numero de paciente es.

Sex: el sexo de la persona.

Rojo: El porcentaje de píxeles rojos en la imagen asociada con el caso.

Verde:El porcentaje de píxeles verdes en la imagen asociada con el caso.

Azul:El porcentaje de píxeles azules en la imagen asociada con el caso.

Hb:Nivel de hemoglobina del individuo, medido en gramos por decilitro (g/dL).

Anaemic:variable categorica que nos dice si tiene o no la enfermedad.

Para nuestro analisis no consideraremos ni la variable number ni sex, ya que nuestro principal interes es discriminar si en base a las variables numericas continuas la persona tiene o no la enfermedad.

A continuacion mostramos un encabezado de los datos donde podemos visualizar los primeros 5 de estos.

Tabla 1: Encabezado de los datos

Rojo	Verde	Azul	Hb	Anaemic
43.2555	30.8421	25.9025	6.3	Yes
45.6033	28.1900	26.2067	13.5	No
45.0107	28.9677	26.0215	11.7	No
44.5398	28.9899	26.4703	13.5	No
43.2870	30.6972	26.0158	12.4	No

Como tenemos dos categorias para la variable **Anaemic**. **Yes** y **No** definiremos a **G1** como el grupo de las personas que dieron **negativo** a anemia y **G2** como el grupo que dieron **positivo** a anemia, adicionalmente definimos n_1 como las observaciones asociadas a $G1$ Y n_2 como las observaciones asociadas a $G2$, en base a esto lo primero que hacemos es calcular cuantas personas estan catalogadas en cada una de las clases.

tenemos que $n_1 = 78$ y $n_2 = 26$, A continuación, mostramos el vector de medias para cada una de las categorías de la variable Anaemic, discriminado para los dos grupos G_1 y G_2

Tabla 2: Vectores de Media

Caracteristicas	No	Yes
Rojo	46.22599	43.938730
Verde	28.36965	30.393080
Azul	25.40436	25.668210
Hb	13.25769	8.830769

Podemos observar que las medias para las variables **Rojo**, **Azul** y **Verde** son bastante similares entre ambos grupos. Sin embargo, se observa una gran diferencia en las medias de la variable **Hb** (Hemoglobina). Esto sugiere que Hb podría ser una variable de interés significativa a la hora de determinar si una persona padece Anemia.

A continuación, mostramos las matrices de varianzas y covarianzas para ambos grupos.

o primero que podemos observar es que las varianzas para las personas que no tienen la enfermedad son ligeramente mayores para las variables **Rojo**, **Azul** y **Verde**. En contraste, para la variable **Hb** (Hemoglobina), las personas con la enfermedad presentan una mayor variabilidad en comparación con las que no la tienen. Esto sugiere que la variable Hb podría tener una incidencia significativa en la probabilidad de dar positivo para el padecimiento de anemia.

Tabla 3: Matriz de Varianzas y Covarianzas para Yes

	Rojo	Verde	Azul	Hb
Rojo	6.5243	-2.0524	-4.4719	2.2084
Verde	-2.0524	1.2689	0.7835	-0.8475
Azul	-4.4719	0.7835	3.6884	-1.3608
Hb	2.2084	-0.8475	-1.3608	4.4990

Tabla 4: Matriz de Varianzas y Covarianzas para No

	Rojo	Verde	Azul	Hb
Rojo	7.1665	-2.4477	-4.7188	0.7383
Verde	-2.4477	1.7826	0.6651	-0.7862
Azul	-4.7188	0.6651	4.0538	0.0480
Hb	0.7383	-0.7862	0.0480	2.0235

Adicionalmente graficamos los dos grupos G_1 y G_2, donde los puntos de color azul indica que no tienen la enfermedad y los de color rojo las que si la tienen.

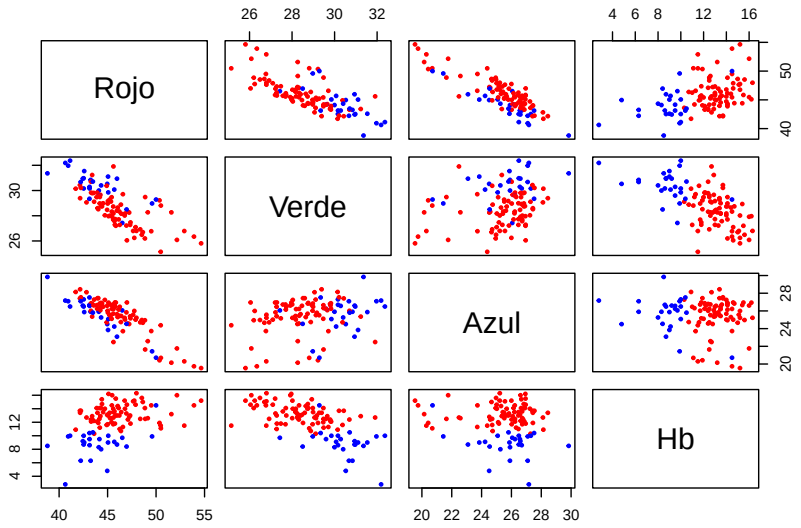

Figura 1: Graficos de dispersion discriminado por resultado

En varios de los gráficos, se observa una discriminación notable entre los dos grupos. Por ejemplo, en la relación entre “Rojo” y “Verde”, así como entre “Rojo” y “Hb”, hay diferencias claras en las posiciones de los puntos rojos y azules. Los puntos rojos (que representan casos negativos de anemia) tienden a agruparse en zonas específicas, mientras que los puntos azules

(que representan casos positivos de anemia) parecen ubicarse en otras regiones, aunque con cierta superposición en algunas relaciones.

Aunque se aprecian diferencias notables en varios gráficos, en algunas relaciones los puntos rojos y azules muestran una mayor superposición. Esto sugiere que esas variables podrían no ser tan efectivas para discriminar entre los dos grupos como otras variables que presentan una separación más clara.

En este punto, tenemos dos posibles caminos para realizar un análisis discriminante: el Análisis Discriminante Lineal (**LDA**) y el Análisis Discriminante Cuadrático (**QDA**). Para determinar cuál de estos métodos usar, debemos evaluar si nuestros datos $\underline{\mathbf{x}}_j \sim N_4(\mu_j, \mathbf{\Sigma})$, $j = 1, 2$, es decir, si las poblaciones de personas que dieron positivo o negativo siguen una distribución normal multivariada.

2 Analisis de Normalidad uni variada

Lo primero para comprobar normalidad es abordarlo de manera grafica, para ello se grafican los respectivos histogramas, para cada grupo G_1 y G_2 y para cada una de las 4 variables.

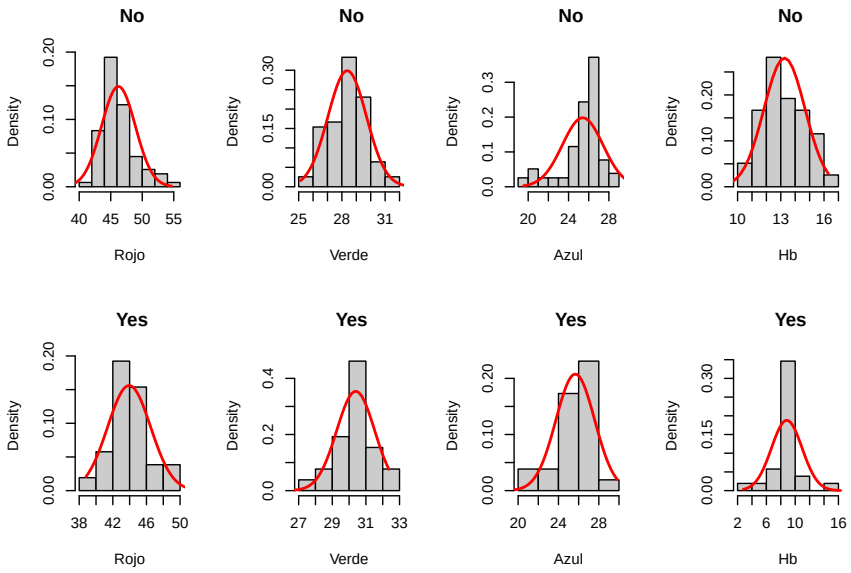

Figura 2: Histogramas por resultado

Lo primero que se observa es que cada uno de ellos tiene una forma acampanada, lo que sugiere que podrían seguir una distribución normal. Sin embargo, esta observación no es prueba suficiente para confirmar este comportamiento. Para hacerlo, realizaremos un gráfico adicional, el QQ-plot, y para ser más precisos, aplicaremos un test estadístico para confirmar la normalidad de la distribución.

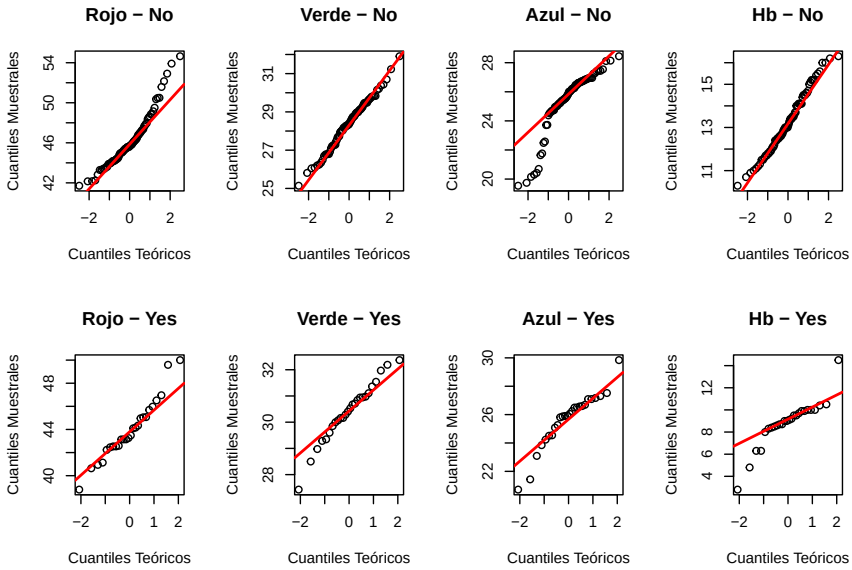

Figura 3: qqplot variable y tipo de resultado

Como podemos observar en los gráficos de normalidad QQ-Plot, hay tres casos en los que los datos no se ajustan a la línea recta: Rojo-NO, Azul-NO y Hb-Yes. Por lo tanto, estos gráficos sugieren que las variables correspondientes no siguen una distribución normal univariada.

Tabla 5: Prueba de normalidad por clasificación

Variable	Estadistico_Shapirto	Valor_p
RojoNo	0.9325554	0.0004707
VerdeNo	0.9931850	0.9561569
AzulNo	0.8498267	0.0000002
HbNo	0.9841708	0.4458559
RojoYes	0.9601808	0.3951046
VerdeYes	0.9746358	0.7448000
AzulYes	0.9244572	0.0572469
HbYes	0.8687570	0.0033616

Para ser más concluyentes, realizamos el test de Shapiro-Wilk, y observamos que el valor p para las tres categorías Rojo-NO, Azul-NO y Hb-Yes es menor a 0.05. Esto nos permite concluir que los datos para estas categorías no siguen una distribución normal univariada. Dado que los datos no son normales univariados, es improbable que lo sean a nivel multivariado, lo que presenta un serio problema ya que los métodos **LDA** y **QDA** requieren la normalidad multivariada.

No obstante, notamos que las otras variables sí cumplen el criterio de normalidad. Por lo tanto, procederemos con el análisis, pero con la precaución de considerar esta limitación.

3 Test M de Box varianzas iguales

Una parte crucial del analisis de discriminantes es saber como son las estructuras de las matrices de varianzas y covarianzas, ya que si estas matrices son iguales para ambos grupos G_1 y G_2 se procede a realizar el analisis en base a el metodo **LDA**, mientras que si esto no ocurre es necesario usar un metodo mas robusto como lo es **QDA**, para ello proponemos la siguiente prueba de hipotesis para los dos grupos poblacionales G_1 y G_2 con matrices de varianzas y covarianzas Σ_1 y Σ_2.

Por lo tanto proponemos el siguiente juego de hipotesis:

$$H_0 : \Sigma_1 = \Sigma_2$$
$$H_a : \Sigma_1 \neq \Sigma_2$$

Usaramos el criterio M de Box donde si obtenemos un valor p mayor a 0.05, concluimos que las matrices de varianzas y covarianzas para ambos grupos son iguales

Tabla 6: Test de M box para Varianza

Estadístico	Valor
Chi-Sq (approx.)	15.3780
df	10.0000
p-value	0.1189

El test de Shapiro-Wilk nos da un valor p de 0.1189, que es mayor a 0.05. Por lo tanto, no podemos rechazar la hipótesis nula de normalidad para esta categoría. Dado que la estructura de covarianzas entre los grupos es igual, podemos aplicar el método de discriminante lineal (LDA).

4 Analisis Mediante discriminante lineal LDA

El Análisis Discriminante Lineal (LDA) es una técnica estadística utilizada ampliamente en la clasificación de datos. Su principal objetivo es encontrar una combinación lineal de características que mejor separe las diferentes categorías en un conjunto de datos. Esta técnica es particularmente eficaz cuando se desea asignar nuevas observaciones a una de varias categorías basándose en características medidas.

En problemas de clasificación, **LDA** busca identificar las características que proporcionan la mejor separación entre clases. Esto se logra mediante la maximización de la distancia entre las medias de las clases mientras se minimiza la variabilidad dentro de cada clase. En términos prácticos, LDA transforma las características originales en un nuevo espacio donde las clases se vuelven más separables, facilitando así la tarea de clasificación.

$$\delta_k(\mathbf{x}) = \mathbf{x}^T \mathbf{W}_k + b_k$$

donde:

- $\mathbf{x}$ es el vector de características.
- $\mathbf{W}_k$ es el vector de pesos discriminantes para la clase k.
- b_k es el término de sesgo para la clase k.

Los vectores de pesos discriminantes $\mathbf{W}_k$ se calculan usando la siguiente fórmula:

$$\mathbf{W}_k = \mathbf{S}_w^{-1}(\mu_k - \mu)$$

donde:

- $\mathbf{S}_w$ es la matriz de varianza-covarianza dentro de las clases (también llamada matriz de varianza-covarianza agrupada).
- μ_k es el vector de medias de la clase k.
- μ es el vector de medias globales.

El término de sesgo b_k se calcula como:

$$b_k = -\frac{1}{2}\mu_k^T \mathbf{S}_w^{-1} \mu_k + \log(\pi_k)$$

donde π_k es la proporción de la clase k en el conjunto de datos.

Como se observó anteriormente, el método LDA es aplicable en este caso. Dado que la estructura de covarianzas entre los grupos es igual y hemos comprobado que los datos cumplen con el criterio de normalidad en al menos una de las categorías, propondremos una solución basada en el método LDA.

En la siguiente tabla se presentan los diferentes casos de asignación de una observación $\underline{x}$ a uno de los dos grupos definidos previamente. En las celdas de la tabla se muestran las asignaciones correspondientes para la clasificación de anemia:

	Decisión	Estadística
Población	G_1	G_2
G_1	Decisión Correcta	Error
G_2	Error	Decisión Correcta

En esta podemos notar cuando una persona esta bien clasificada o mal clasificada segun el algoritmo.

Tabla 7: Matriz de confusión del modelo LDA

Anemia	NO	Yes
No	78	3
Yes	0	23

Podemos notar que el método LDA clasifica incorrectamente a 3 personas que no tenían anemia, pero el método las clasifica como positivas para la enfermedad. Por lo tanto, definimos la tasa de error (TEA) como:

$$TEA = \frac{3+0}{78+3+0+23} = 0.0288$$

La tasa de error (TEA) del modelo es de 2.88%. Este valor es muy bajo, lo que indica que el método LDA discrimina de manera muy efectiva y clasifica a las personas de manera adecuada.

5 Conclusiones

El Análisis Discriminante Lineal (LDA) es una herramienta poderosa para la clasificación que permite identificar las características más relevantes para separar diferentes categorías. Su capacidad para reducir la dimensionalidad y mejorar la separación entre clases lo convierte en una técnica valiosa en el análisis de datos y la modelización predictiva.

El LDA es más eficaz cuando los datos cumplen con ciertas condiciones, como la normalidad multivariada. Aunque este es un requisito, en la práctica, el LDA puede ser bastante efectivo siempre que las matrices de varianzas y covarianzas para los distintos grupos sean iguales. La igualdad de las matrices de covarianzas es crucial, ya que asegura que el modelo de LDA pueda realizar una separación óptima entre las clases.

Referencias

Jhonson, D., Richard And Wichern. (2007). *Applications of Multivariate Technique*. Pearson Education.

Luque-Calvo, P.L. (2017). *Escribir un Trabajo Fin de Estudios con R Markdown*. Disponible en http://destio.us.es/calvo.

Porras, J.C. (2016). Comparacion de pruebas de normalidad multivariada. *Anales Cientificos*, pp. 141-146. Universidad Nacional Agraria La Molina.

Royston, P. (1992). Approximating the Shapiro-Wilk W-test for non-normality. *Statistics and computing*, **2**, 117-119.