Godinez Bravo Diego

**Estadística Multivariada**

Tarea 3

CIMAT

**Ejercicio 1.** Considera los datos de Forbes dados el archivo P1-4.DAT que corresponde a los datos de compañías mas grandes de los USA en 1990.
Realice lo siguiente:

- Ajuste un modelo de regresión lineal a los datos, considerando profits como la variable dependiente y sales y assets como la variable independiente.
- Determine los intervalos de confianza simultáneos e individuales para un nivel de significancia de $\alpha = 0{,}05$.
- Analice los residuales y discuta si el modelo es adecuado. Calcule los leverage points asociados. ¿Algunas de estas compañías pueden considerarse datos atípicos en el conjunto de las variables explicativas?
- Interprete sus respuestas.

**Solución**

Coeficientes obtenidos a partir del análisis de regresión para variable de respuesta $y =$ profits.

| | |
|---|---|
| Intercepto | 0.01332 |
| Coeficiente $\beta_1$ | 0.06805 |
| Coeficiente $\beta_2$ | 0.00576 |

**Solución**

Recordamos que los intervalos de confianza simultáneos del $100(1-\alpha)\,\%$ para las $\beta_i$ están dados por

$$\hat{\beta}_i \pm \sqrt{\hat{var}(\hat{\beta}_i)}\sqrt{(r+1)F_{r+1,n-r-1}\,(\alpha)} \qquad i = 0, 1, ..., r$$

donde $\sqrt{\hat{var}(\hat{\beta}_i)}$ es el i-esimo elemento de la diagonal de $s^2(Z'Z)^{-1}$ correspondientes a $\hat{\beta}_i$

Con frecuencia, en la práctica se ignora la propiedad de confianza simultánea en las estimaciones de los intervalos para los coeficientes $\beta_i$. Entonces se reemplaza el término $(r+1)F_{r+1,n-r-1}\,(\alpha)$ por el $t$ valor $t_{n-r-1}(\alpha)$, que es el percentil de una $t$ asumiendo que los intervalos para cada $\beta_i$ se obtienen de manera univariada.De manera que los intervalos individuales se encuentran definidos de la siguiente manera

$$\hat{\beta}_i \pm \sqrt{\hat{var}(\hat{\beta}_i)}\sqrt{(r+1)t_{n-r-1}(\alpha)}$$

Haciendo uso del lenguaje de programación R se obtuvo el cálculo de los intervalos de confianza simultáneos e individuales.

Intervalos de Confianza Simultáneos

| Limite Inferior | $\beta_i$ | Limite Superior |
|---|---|---|
| -27.581 | $\leq \beta_0 \leq$ | 27.607 |
| -0.0325 | $\leq \beta_1 \leq$ | 0.1686 |
| -0.0120 | $\leq \beta_2 \leq$ | 0.0236 |

Intervalos de Confianza Individuales

| Limite Inferior | $\beta_i$ | Limite Superior |
|---|---|---|
| -11.737 | $\leq \beta_0 \leq$ | 11.763 |
| 0.0252 | $\leq \beta_1 \leq$ | 0.1108 |
| -0.0018 | $\leq \beta_2 \leq$ | 0.0133 |

**Solución**

Basándonos en los gráficos obtenidos del análisis de residuales, no podríamos sugerir que el modelo está ajustado correctamente. A pesar de que se observa que los residuos no muestran un patrón aparente, a simple vista, podemos observar valores atípicos en el gráfico QQ que podrían refutar el supuesto de una normalidad, es necesario confirmar la presencia de datos atípicos generando intervalos de confianza adecuados para datos rsiduales y confirmar o refutar el supuesto de normalidad (**Figura 1**).

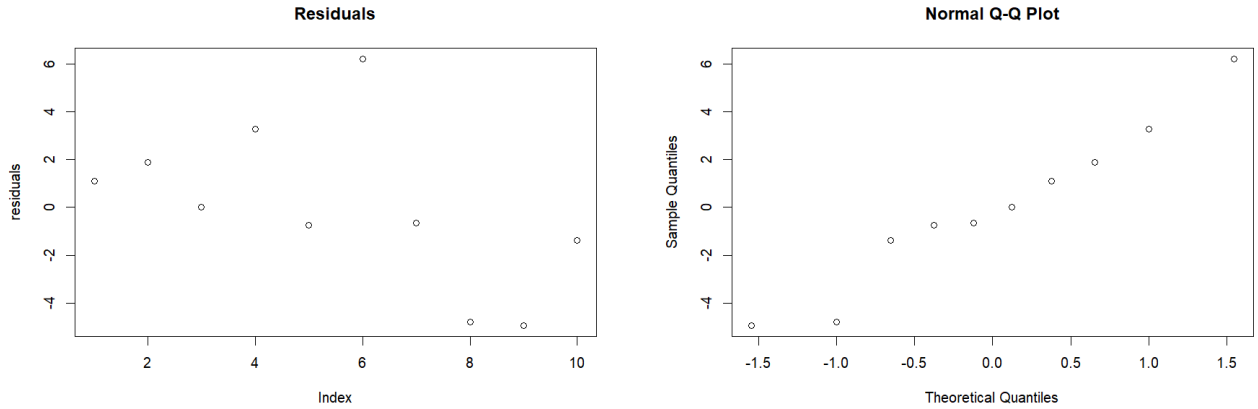


**Figura 1.** Análisis de regresión considerando para la variable de respuesta $y =$ profits.

**Ejercicio 2.** Considere los datos de contaminantes dados en el archivo T1-5.DAT que corresponde a los datos de contaminantes a mediodía en los Angeles. Sean $y_1 = NO_2$ y $y_2 = O_3$ las dos respuestas (contaminantes) correspondientes a las variables predictoras $z_1 =$ wind y $z_2 =$ solar radiation.

- Realice un análisis de regresión utilizando solamente la primera respuesta $y_1$.
  - Sugiera y ajuste un modelo de regresión lineal apropiado.
  - Analice los residuales.
- Realice un análisis de regresión multivariado utilizando ambas respuestas $y_1$ y $y_2$.
  - Sugiera y ajuste un modelo de regresión lineal apropiado.
  - Analice los residuales.

**Solución**

Coeficientes obtenidos a partir del análisis de regresión para variable de respuesta $y_1 = NO_2$.

| | |
|---|---|
| Intercepto | 10.1145 |
| Coeficiente $\beta_1$ | -0.2112 |
| Coeficiente $\beta_2$ | 0.0205 |

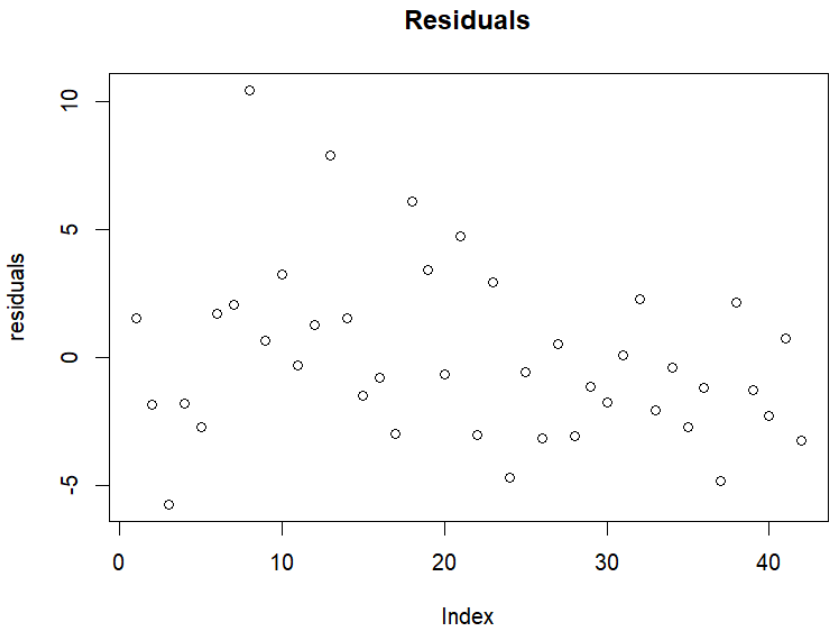


**Figura 2.** Análisis de residuales obtenidos a partir del análisis de regresión para $y_1$.

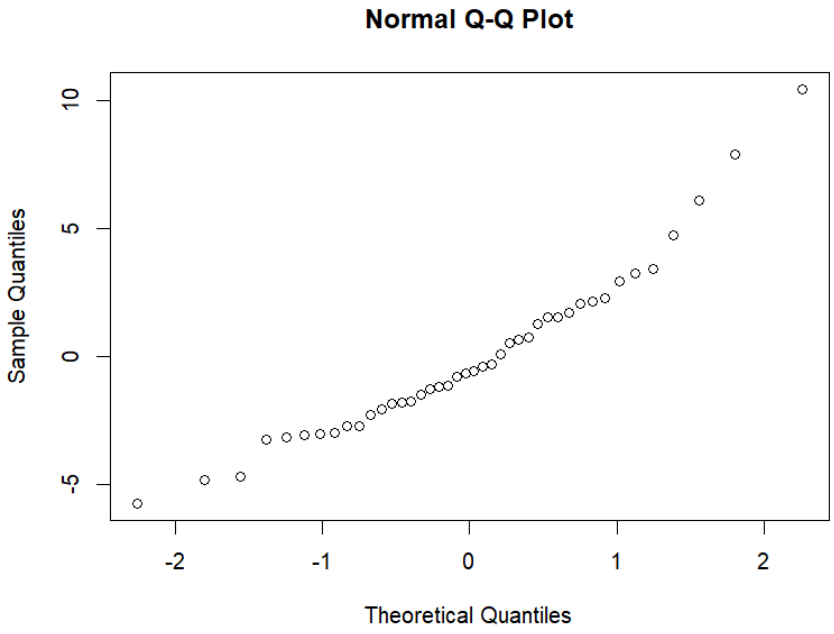


**Figura 3.** Gráfico QQ plot de residuales obtenidos a partir del análisis de regresión para $y_1$

Basándonos en los gráficos utilizados para analizar la distribución de los residuales, podemos concluir que el modelo está ajustado correctamente. Esto se evidencia ya que los residuos no muestran un patrón aparente (**Figura 2**) y, a simple vista, exhiben una distribución normal (**Figura 3**). Sin embargo, es importante señalar que para validar esta suposición de normalidad, sería necesario generar intervalos de confianza que refuercen lo establecido previamente.

**Solución**

Cuando consideramos la extensión multivariada de la regresión lineal multiple, se tiene la modelación de $m$ respuestas $Y_1, Y_2, ..., Y_m$ y un conjunto de $r$ variables predictoras $z_1, z_2, ..., z_r$.

$$Y_1 = \beta_{01} + \beta_{11} z_1 + ... + \beta_{r1} z_r + \epsilon_1$$

$$Y_2 = \beta_{02} + \beta_{12} z_1 + ... + \beta_{r2} z_r + \epsilon_2$$

$$\vdots$$

$$Y_m = \beta_{0m} + \beta_{1m} z_1 + ... + \beta_{rm} z_r + \epsilon_m$$

De manera que se asume que **cada una de las respuestas sigue su propio modelo de regresión** cada una con distintos coeficientes $\beta$ pero **con la misma matriz de diseño Z**.

Considerando lo anterior, se realizó un análisis de regresión por cada variable de respuesta.

Coeficientes obtenidos a partir del análisis de regresión para variable de respuesta $y_1 = NO_2$.

| | |
|---|---|
| Intercepto | 10.1145 |
| Coeficiente $\beta_1$ | -0.2112 |
| Coeficiente $\beta_2$ | 0.0205 |

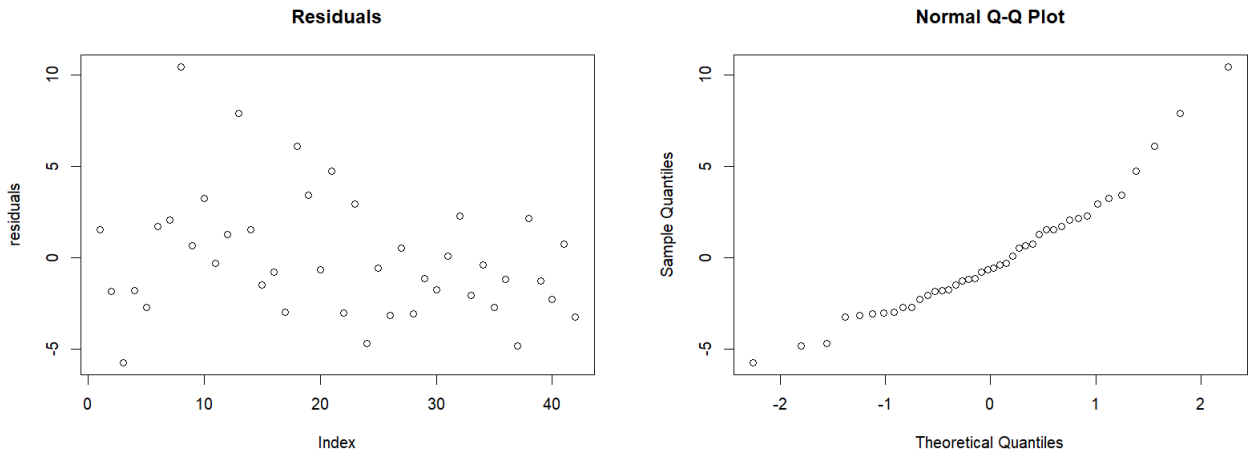


**Figura 4.** Análisis de regresión considerando la variable de respuesta $y_1$.

Coeficientes obtenidos a partir del análisis de regresión para variable de respuesta $y_2 = O_3$.

| | |
|---|---|
| Intercepto | 8.2761 |
| Coeficiente $\beta_1$ | -0.7868 |
| Coeficiente $\beta_2$ | 0.0951 |

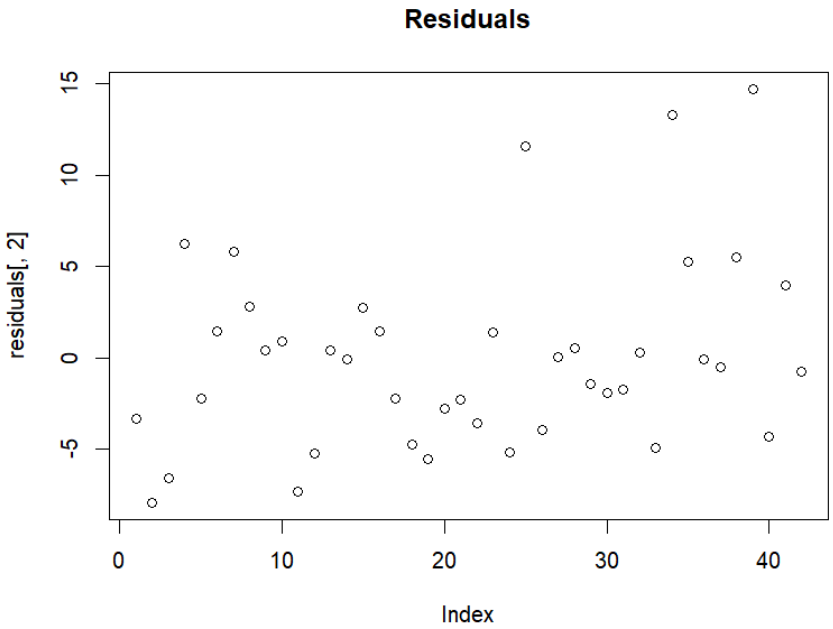


**Figura 5.** Análisis de residuales obtenidos a partir del análisis de regresión para $y_2$.

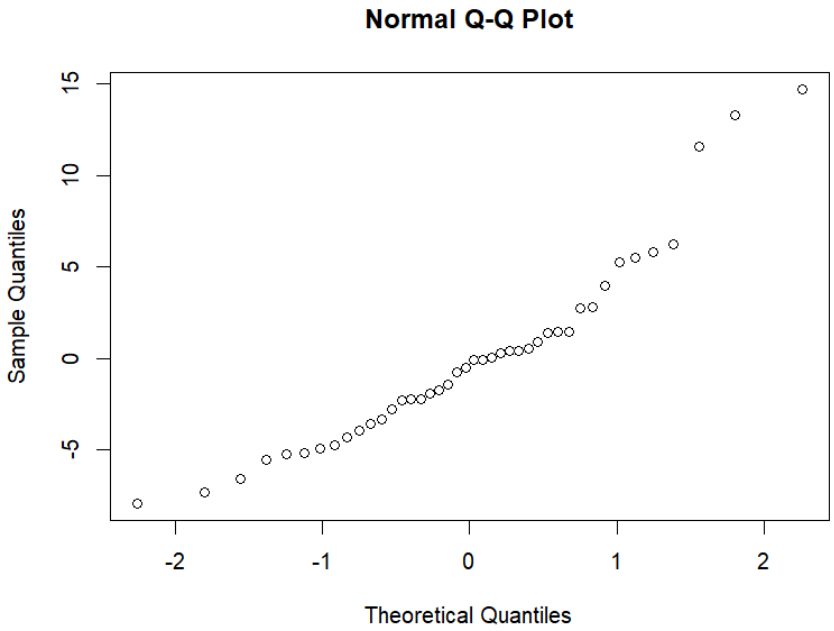


**Figura 6.** Gráfico QQ plot de residuales obtenidos a partir del análisis de regresión para $y_2$.

Al igual que en el caso anterior, podemos observar que los residuales no muestran un patrón evidente (**Figura 5**), no obstante, se observa una pequeña desviación en las colas en el gráfico QQ (**Figura 6**). Por lo tanto, es necesario evaluar el supuesto de normalidad mediante la generación de intervalos de confianza adecuados que nos permitan corroborar su distribución.

**Ejercicio 3.** En momentos de estrés, el cuerpo sufre un proceso llamado vasoconstricción en el que los vasos sanguíneos de las extremidades se cierran, forzando la sangre a los órganos centrales. La vasoconstricción también puede ocurrir después de respirar profundamente. El conjunto de datos **vaso** en la biblioteca *robustbase* resume la vasoconstricción (o no) de los dedos de los sujetos junto con sus volúmenes y frecuencias respiratorios.

a) Grafique las tasas y volúmenes respiratorios usando diferentes colores para aquellos con y sin vasoconstricción. Realice un análisis discriminante y vea qué tan bien este distingue entre los dos grupos.

b) Repita el análisis de la parte (a), pero primero aplique logaritmos a las variables explicativas. Compare ambos análisis.

**Solución**

El objetivo del Análisis Lineal Discriminante (LDA) es clasificar objetos en uno o más grupos basados en un conjunto de características que describen los objetos.

Los coeficientes de las funciones discriminantes son útiles para describir diferencias grupales e identificar variables que distinguen entre grupos. Representan la contribución de cada variable a la clasificación de los grupos en el espacio de las funciones discriminantes.

Coeficientes obtenidos de la función discriminante LD1.

| | LD1 |
|---|---|
| Volume | 1.5713 |
| Rate | 1.3453 |

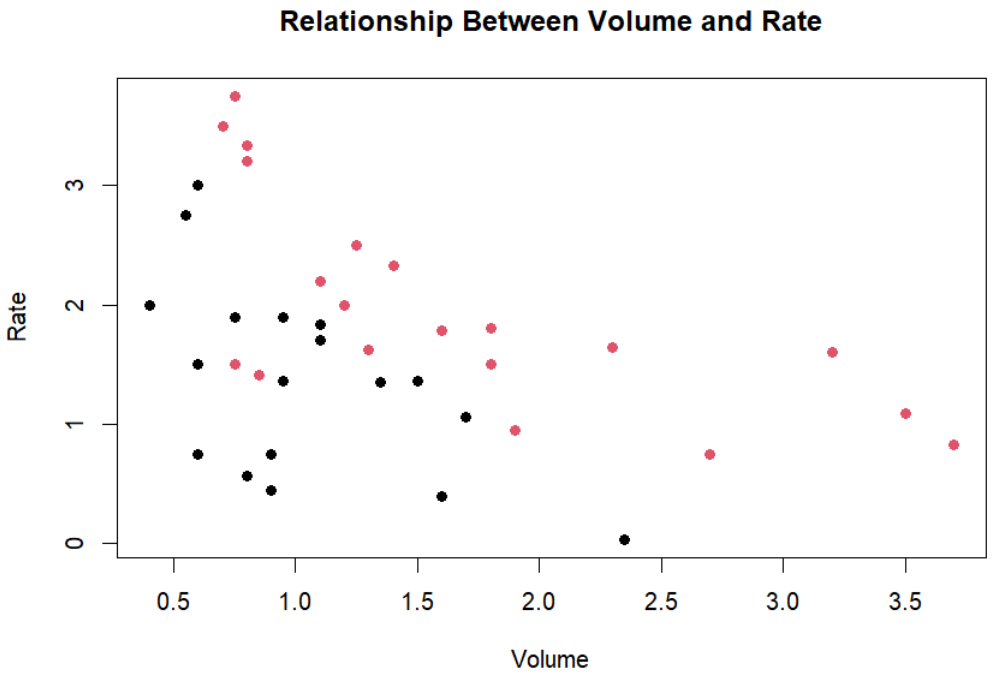


**Figura 7.** Relación entre variables *Volumen* y *Tasa*.

**Solución**

Al aplicar el logaritmo a ambas variables explicativas (i.e. *Volume*, *Rate*), se obtuvieron los siguientes coeficientes de la función discriminante LD1.

| | LD1 |
|---|---|
| Volume | 2.2683 |
| Rate | 1.3262 |

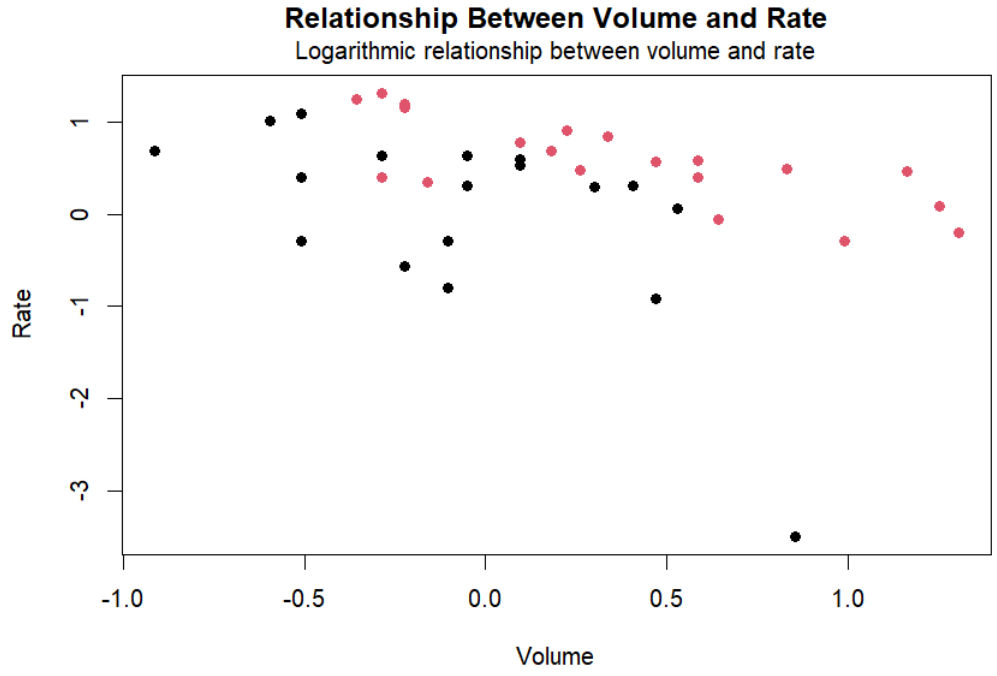


**Figura 8.** Relación entre variables *Volumen* y *Tasa* aplicando el logaritmo en ambas variables.

Observamos que la relación entre las variables se mantiene al aplicar el logaritmo, donde el aumento de volumen influye en la vasoconstricción (**Figura 7**; **Figura 8**).

De acuerdo con los coeficientes obtenidos de la función discriminante **LD1**, podemos sugerir que el volumen contribuye en mayor medida a la diferenciación de ambos grupos. Esto se confirma al aplicar el modelo de análisis discriminante a las variables transformadas, donde la diferencia entre los coeficientes es mayor, siendo el volumen el factor que más contribuye a la separación de las clases. Este resultado coincide con lo mencionado anteriormente.

**Ejercicio 4.** En el archivo wine.dat se encuentran los datos de 178 vinos.

a) Utilice las variables descriptoras para crear un discriminante lineal para clasificar a cada vino de acuerdo del cultivo de uva que proviene.

b) Grafique los vinos en el espacio discriminante propuesto.

c) Verifique la precisión de la regla discriminante propuesta, utilizando los métodos de clasificación cruzada y conjuntos de entrenamiento-prueba.

**Solución**

El objetivo del Análisis Lineal Discriminante (LDA) es clasificar objetos en uno o más grupos basados en un conjunto de características que describen los objetos.

Los coeficientes de las funciones discriminantes son útiles para describir diferencias grupales e identificar variables que distinguen entre grupos. Representan la contribución de cada variable a la clasificación de los grupos en el espacio de las funciones discriminantes.

A partir del análisis discriminante se obtuvieron los siguientes coeficientes para las funciones discriminantes

| | LD1 | LD2 |
|---|---|---|
| alcohol | -0.4034 | 0.8718 |
| malic | 0.1653 | 0.3054 |
| ash | -0.3691 | 2.3458 |
| alcal | 0.1548 | -0.1464 |
| mg | -0.0022 | -0.0005 |
| phenol | 0.6181 | -0.0322 |
| flav | -1.6612 | -0.4920 |
| nonf | -1.4958 | -1.6310 |
| proan | 0.1341 | -0.3071 |
| color | 0.3551 | 0.2532 |
| hue | -0.8180 | -1.5156 |
| abs | -1.1576 | 0.0512 |
| proline | -0.0027 | 0.0029 |

Considerando lo mencionado previamente, podemos afirmar que las variables que contribuyen en mayor medida a la distinción de las clases de vinos son **flav**, **nonf**, **abs**, **hue** y **phenol**; esto para la dimensión LD1. Por otro lado, para la dimensión LD2, las variables **ash** y **alcohol**, tienen un coeficiente alto en comparación con otras variables, lo que resalta su relevancia en esta dimensión.

De manera general, al interpretar estos coeficientes podemos comprender mejor qué caracteríísticas de los vinos contribuyen en mayor medida a la discriminación entre las clases de vinos en el análisis de LDA.

**Solución**

Observamos una clara distinción entre las tres clases de vinos, lo que sugiere que las variables seleccionadas contribuyen de manera significativa para distinguir entre estas categorías. Cada clase muestra una agrupación compacta y bien definida, lo que indica que las variables utilizadas en el análisis proporcionan una discriminación robusta entre ellas (**Figura 9**).

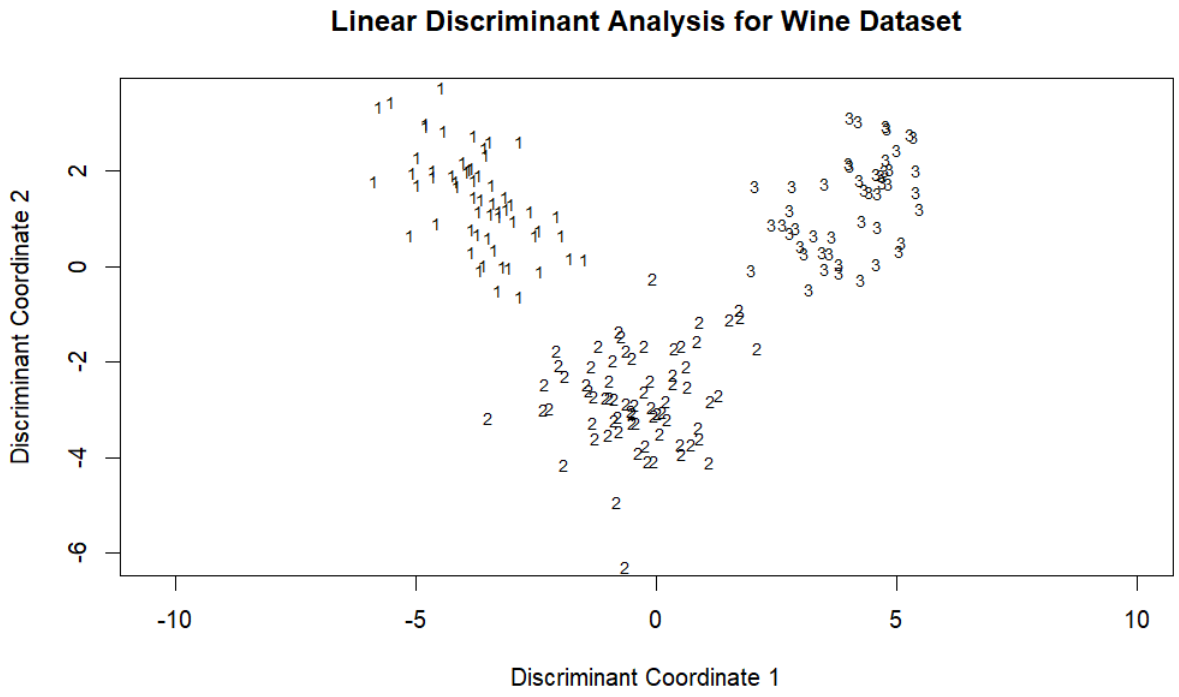


**Figura 9.** Análisis Lineal Discriminante para el conjunto de datos wine.

**Solución**

Basándonos en el valor de accuracy obtenido podemos afirmar que el modelo es altamente preciso en la clasificación de clases de vinos definidas en el conjunto de datos. Esto indica que el modelo captura eficazmente los patrones presentes en las características de los vinos y puede distinguir entre las diferentes clases con gran precisión.

| Accuracy | Kappa |
|---|---|
| 0.9944 | 0.9916 |