Regresión Lineal II

Probabilidad Aplicada

3602

Análisis de regresión lineal: varias variables simultáneas

En este capítulo veremos cómo gestionar un análisis de regresión lineal a partir de una base que cuenta con muchas variables de múltiples de diferentes tipos.

Utilizaremos la base **mtcars** de R base y las librerías…

En primer lugar, cargaremos la base:

```
autos <- mtcars
```

Esta base o conjunto de datos (datset) incluye información sobre diferentes modelos de automóviles. Contiene 32 observaciones (automóviles) y 11 variables numéricas que representan diversas características del rendimiento y especificaciones técnicas de los vehículos.

Tabla de variables

Variable	Representa	Tipo de variable
mpg	Millas por galón (Miles per Gallon). Representa el rendimiento de combustible.	Continua
cyl	Cilindros. Indica el número de cilindros en el motor del automóvil	Discreta
disp	Desplazamiento del motor (Displacement). Volumen total de los cilindros del motor medido en pulgadas cúbicas (cubic inches).	Continua
hp	Caballos de fuerza (Horsepower). Potencia del motor del automóvil medida en caballos de fuerza.	Continua
drat	Relación del eje trasero (Rear Axle Ratio). Es la relación de transmisión del eje trasero, lo que afecta la velocidad y la eficiencia del motor.	Continua
wt	Peso del vehículo (Weight). Peso del automóvil en miles de libras.	Continua
qsec	Tiempo en recorrer un cuarto de milla (1/4 mile time).	Continua
vs	Disposición del motor (Engine Shape). Tipo de configuración del motor: 0 = Motor en disposición de cilindros en línea (V-shaped engine). 1 = Motor en línea recta (Straight engine).	Categórica discretizada.
am	Tipo de transmisión (Transmission). Tipo de transmisión del automóvil: 0 = Transmisión automática. 1 = Transmisión manual.	Categórica discretizada.
gear	Número de marchas (Number of Forward Gears). Cantidad de marchas hacia adelante con las que cuenta el automóvil (3, 4 o 5 marchas).	Discreta
carb	Número de carburadores (Number of Carburetors).	Discreta

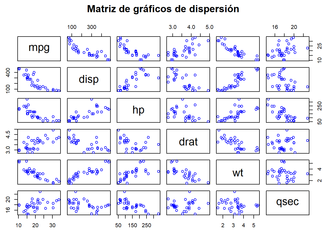
Como deseamos correlacionar variables con el método de regresión lineal, sólo estudiaremos las variables continuas. Entonces hacemos un select() de las variables de interés.

```
autos <- autos %>%
    select(mpg, disp, hp, drat, wt, qsec)
```

Matriz de gráficos de dispersión

Una vez actualizada nuestra base con las variables de interés, realizaremos todas las combinaciones de variables en una **matriz de gráficas de dispersion**. Para ello, utilizaremos la función **pairs()** de R base.

```
pairs(autos, #base
      col="blue", #color de la dispersion
      main = "Matriz de gráficos de dispersión" #título
      )
```



Interpretación de la matriz gráfica de dispersión

Gráficos de dispersión

Cada gráfico muestra la relación entre un par de variables. Por ejemplo, el gráfico en la intersección de la fila **mpg** (Millas por Galón) y la columna **hp** (Caballos de Fuerza) muestra cómo el consumo de combustible se relaciona con la potencia del motor.

La forma de la nube de puntos sugiere el tipo de relación. Si hay una relación lineal, cuadrática u otra relación entre las variables.

Si los puntos tienden a alinearse en una dirección (ascendente o descendente), es probable que haya una relación lineal.

Diagonal principal:

En la diagonal principal de la matriz (desde la esquina superior izquierda hasta la esquina inferior derecha), normalmente se muestran los nombres de las variables.

Relaciones Positivas

Si los puntos en un gráfico de dispersión muestran una tendencia ascendente, es decir, cuando una variable aumenta, la otra también tiende a aumentar, esto indica una correlación positiva.

Ejemplo: Si los gráficos entre **wt** (peso) y **disp** (desplazamiento) muestran una tendencia ascendente, esto indica que los automóviles con mayor peso tienden a tener motores de mayor desplazamiento.

Relaciones Negativas

Si los puntos en un gráfico de dispersión muestran una tendencia descendente, es decir, cuando una variable aumenta, la otra disminuye, esto indica una correlación negativa.

Ejemplo: A mayor peso (**wt**), menor es la eficiencia en términos de millas por galón (**mpg**). Una nube descendente en esta celda indicaría esta relación.

Relaciones Débiles

Si los puntos están dispersos sin una tendencia clara (ni ascendente ni descendente), indica que hay una correlación débil entre esas dos variables.

Simetría

Los gráficos de dispersión son simétricos a través de la diagonal. Por ejemplo, el gráfico en la celda de la fila **mpg** y la columna **hp** es el mismo que el gráfico en la fila **hp** y la columna **mpg**. Solo cambia el orden de las variables en los ejes x e y.

Matriz de correlación (corplot)

Una utilidad de la matriz de dispersión es que nos da indicios para seleccionar los pares de variables que pueden relacionarse linealmente. Pero, ¿qué tan buena es nuestra primera selección? Necesitamos más que un indicio. Vamos a calcular los coeficientes de correlación r para cada par de variables combinadas y la dispondremos en una matriz.

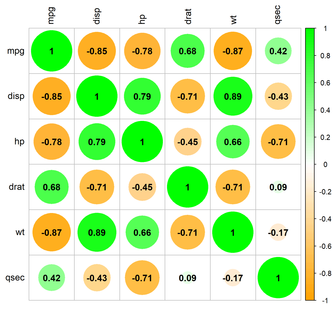
Para ello, utilizaremos la función **cor()** y **corrplot()** del paquete **corrplot**

En primer lugar, creamos la matriz de correlación en el objeto **matriz_cor**

```
matriz_cor <- cor(autos)
```

Luego le aplicamos la función **corrplot()** a nuestra matriz de correlación para obtener el gráfico "corplot":

```
corrplot(matriz_cor,
         method = "circle",     # Representa las correlaciones con círculos
         type = "full",         # Muestra la matriz completa (ambos lados simétricos)
         title = "Matriz de Correlación", # Título
         cex.main = 1, #Tamaño del título
         tl.col = "black",      # Color de las etiquetas de las variables
         addCoef.col = "black", # Mostrar los valores numéricos de las correlaciones
         col = colorRampPalette(c("orange", "white", "green"))(200),# Colores
         mar = c(0,0,2,0))  #Ajuste de márgenes
```



Interpretación de la matriz de correlación

Valores de Correlación

La matriz muestra coeficientes de correlación que van desde -1 hasta 1.

Valores cercanos a 1: Indican una correlación positiva fuerte. Esto significa que a medida que una variable aumenta, la otra también tiende a aumentar.

Valores cercanos a -1: Indican una correlación negativa fuerte. Esto significa que a medida que una variable aumenta, la otra tiende a disminuir.

Valores cercanos a 0: Indican que no hay una correlación lineal significativa entre las variables.

Simetría

La matriz es simétrica, lo que significa que la correlación entre la variable A y B es la misma que la correlación entre B y A.

Color

El color de la forma (en este caso círculos) en el gráfico puede variar dependiendo de la dirección y la fuerza de la correlación. Generalmente, los colores más oscuros indican correlaciones más fuertes, mientras que los colores más claros indican correlaciones más débiles.

Las correlaciones positivas a menudo se representan en un color (como azul), mientras que las correlaciones negativas se representan en otro (como rojo).

En este caso, se personalizó una paleta de colores y las correlaciones positivas están en verde mientras que las negativas están en naranja.

Tamaño del valor y forma

El tamaño del valor también puede reflejar la fuerza de la correlación, al igual que el tamaño de la forma. Un tamaño más grande puede indicar una correlación más fuerte, mientras que un tamaño más pequeño puede indicar una correlación más débil.

Ejemplo de Interpretación

Ejemplo de matriz de correlación

	mpg	hp	wt
mpg	1	-0.78	-0.87
hp	-0.78	1	0.66
wt	-0.87	0.66	1

mpg y hp: Tienen una correlación de -0.78, lo que indica que hay una correlación negativa fuerte entre el rendimiento del combustible y los caballos de fuerza. A medida que aumenta el HP, el MPG tiende a disminuir.

mpg y wt: Tienen una correlación de -0.87, lo que indica una correlación negativa muy fuerte. Esto sugiere que a medida que aumenta el peso del automóvil, el rendimiento del combustible disminuye significativamente.

hp y wt: Tienen una correlación de 0.66, lo que indica una correlación positiva moderada. Esto sugiere que a medida que aumentan los caballos de fuerza, también tiende a aumentar el peso del automóvil.