Uso de los MLG en los casos de asociaciones no linealeales entre las variables

Angie Rodríguez Duque & César Saavedra Vanegas

Octubre de 2020

Introducción

Definición

Hasta ahora solo hemos considerado asociaciones lineales entre X e y, donde un aumento de delta en una variable explicativa continua x_i produce el mismo cambio β_1 en y para todos los valores de xi. β_1 a veces se denomina **"Slope"** porque es un gradiente lineal. Una ecuación de regresión lineal simple con una sola pendiente lineal es:

$$E(Y_i) = \beta_0 + \beta_1 x_i; \qquad i = 1, ..., N.$$

Una asociación en forma de U se puede modelar agregando una versión cuadrática de la variable y un parámetro β adicional:

$$E(Y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2; \qquad i = 1, ..., N.$$

Centrar y escalar

En la práctica, cuando se utilizan transformaciones como la cuadrática, que pueden crear valores grandes de xi, puede resultar útil centrar las variables explicativas utilizando su media (x) y escalarlas utilizando su desviación estándar (de). Para mayor comodidad de notación, primero creamos una versión centrada y escalada de x_i:

$$\tilde{x}_i = \frac{(x_i - \bar{x}_i)}{sd}$$

Tabla: Estimaciones para el modelo utilizando variables explicativas centradas y escaladas.

Término	Estimación b_j	Error estándar
Constante	37.600	1.332
Coeficiente para la edad	-1.452	1.397
Coeficiente de peso	-3.793	1.385
Coeficiente de proteína	4.350	1.411

Ajuste del modelo

Y se ajusta al modelo:

$$E(Y_i) = \beta_0 + \beta_1 \tilde{x}_i + \beta_2 \tilde{x}_i^2$$

Ventajas

- Una ventaja adicional del centrado es que la estimación de la intersección β_0 ahora relaciona el valor de y promedio con el valor de x promedio en lugar del valor de y promedio cuando x es cero, lo que puede no ser significativo si x no puede ser cero. **Ejemplo:** El peso de una persona.
- Además, los parámetros de **"Slope"** ahora representan un cambio de una desviación estándar que es potencialmente más significativo que un cambio de una sola unidad que puede ser muy pequeño o grande.
- Por último, escalar por la desviación estándar facilita la comparación de la importancia de las variables.

Ejemplo en R

Datos de la revista PLOS Medicine

Los datos corresponden a 878 artículos de revistas publicados en la revista PLOS Medicine entre 2011 y 2015. El gráfico muestra el número de autores en el eje x y la longitud del título del artículo (incluidos los espacios) en el eje y. Hubo 15 artículos con más de 30 autores que se truncaron a 30. Como el número de autores es discreto, un diagrama de dispersión estándar probablemente tergiversaría los datos ya que los puntos se superpondrían y, por lo tanto, se ocultarían. Para evitar esto, se alteraron los puntos, lo que significa que se agregó un pequeño valor aleatorio a cada punto para evitar la superposición.

Bibliografía

- Dobson, A. J., & Barnett, A. G. (2018). An introduction to generalized linear models. CRC press.