Análisis y Regresión Lineal 2025

Información del dataset utilizado:

- https://archive.ics.uci.edu/dataset/53/iris

Repositorio con algoritmos aplicados en el trabajo:

- https://github.com/nfacundomendoza/Mate4

Para este análisis elegimos el conjunto de datos Iris (Fisher, 1936).

En biología, particularmente en el estudio morfológico de las flores del género Iris, la longitud de los pétalos constituye un dato de especial interés, ya que representa una característica distintiva entre las especies.

El objetivo de este trabajo es aplicar técnicas de regresión lineal para analizar la relación entre las características morfológicas de las flores del dataset. En particular, se busca identificar la variable predictora más relevante mediante regresión simple, construir un modelo de regresión múltiple, comparar los enfoques de mínimos cuadrados y descenso del gradiente, y evaluar cómo la inclusión de variables adicionales mejora la capacidad predictiva del modelo.

Regresión Lineal Simple

A) En este caso, se selecciona la longitud del pétalo (petal length) como variable respuesta, ya que predecir este dato nos da información sobre la especie de la flor y su morfología general. Además, se trata de un dato continuo y no categórico, que presenta variaciones significativas entre especies. Estas diferencias aportan información discriminante útil para la modelización de regresión lineal y múltiple.

Como variables predictoras se consideran la longitud del sépalo (sepal length), el ancho del sépalo (sepal width) y el ancho del pétalo (petal width).

B) Utilizando el algoritmo “regresión-lineal” comparamos las distintas variables predictoras y obtuvimos los siguientes resultados:

sepal length	
$\hat{Y}_i = \beta_1 x_i + \beta_0$	$1,8575096654214451.x.(sepal\ length) - 7,095381478279317$
σ^2	$0,7523515250621987$
R^2	$0,7599553107783261$
r	$0,8717541573048712$
$IC(\beta_1)$	$(1.6879,\ 2.0271)$

IC(β_0)	$(-\ 8.0961,\ \ -\ 6.0947)$
ICM(Y)	-7.0954 + 1.8575·x* ± t_{α/2,148} · √[0.7524 · (1/150 + (x* − 5.8433)²/S_{xx})]
IP(Y)	-7.0954 + 1.8575·x* ± t_{α/2,148} · √[0.7524 · (1 + 1/150 + (x* − 5.8433)²/S_{xx})]

sepal width	
$\hat{Y}_i = \beta_1 x_i + \beta_0$	$-\ 1,7112013879468533.\ x.\ (sepal\ width)\ \ +\ \ 8,98467570545636$
σ^2	$2,57997940920151$
R^2	$0,1768337873324649$
r	$-\ 0,42051609640115445$
IC(β_1)	$(-\ 2.3109,\ \ -\ 1.1115)$
IC(β_0)	$(7.1349,\ 10.8345)$
ICM(Y)	8.9847 − 1.7112·x* ± t_{α/2,148} · √[2.58 · (1 + 1/150 + (x* − 3.054)²/S_{xx})]
IP(Y)	8.9847 − 1.7112·x* ± t_{α/2,148} · √[2.58 · (1 + 1/150 + (x* − 3.054)²/S_{xx})]

petal width	
$\hat{Y}_i = \beta_1 x_i + \beta_0$	$2,2258853065539115\ .\ (petal\ width)\ \ +\ \ 1,0905721458773807$
σ^2	$0,22910722512285006$
R^2	$0,9269012279220037$
r	$0,962757097050966$
IC(β_1)	$(2.1243,\ 2.3274)$
IC(β_0)	$(0.9464,\ 1.2347)$
ICM(Y)	1.0906 + 2.2259·x* ± t_{α/2,148} · √[0.2291 · (1/150 + (x* − 1.1987)²/S_{xx})]
IP(Y)	1.0906 + 2.2259·x* ± t_{α/2,148} · √[0.2291 · (1 + 1/150 + (x* − 1.1987)²/S_{xx})]

C) Luego de analizar los resultados de cada variable predictora, se concluye que la mejor variable es el ancho del pétalo (petal width), por las siguientes razones:

- **Coeficiente de determinación (R^2 = 0,9269):** indica que el 92,69 % de la variabilidad observada en la longitud del pétalo es explicada por el ancho del pétalo, mostrando un excelente ajuste del modelo.

- **Varianza de los residuos (σ^2 = 0,2291):** al ser baja, refleja que las predicciones del modelo se aproximan bastante a los valores observados.

- **Intervalo de confianza para β_1 (2,1243; 2,3274):** dado que no incluye el valor 0, se concluye que el coeficiente es estadísticamente significativo, mostrando una relación lineal positiva entre el ancho y la longitud del pétalo.

- **Gráfico de dispersión:** los valores del conjunto de datos muestran una tendencia lineal clara.

La ecuación del modelo resultante es:

$$\hat{y} = 2,2258853065539115 \cdot (petal\ width) + 1,0905721458773807$$

Donde ŷ es la longitud estimada del pétalo (petal length). Esto indica que, en promedio, por cada incremento de 1 cm en el ancho del pétalo, la longitud del pétalo aumenta aproximadamente 2,23 cm, manteniéndose una relación lineal positiva entre ambas variables.
Por otro lado, en la interpretación biológica, la fuerte correlación entre ancho y largo del pétalo refleja que ambas medidas corresponden a la misma estructura floral, la cual crece de manera proporcional.

Regresión Lineal Múltiple

D) Utilizando el algoritmo “descenso-gradiente” estimamos la ecuación de regresión múltiple:

$$-0.4126 + (0.7937 \cdot sepal\_length) + (-0.6909 \cdot sepal\_width) + (1.3707 \cdot petal\_width)$$

E) Utilizando el algoritmo “minimos-cuadrados” estimamos la ecuación de regresión múltiple:

$$-0.2527 + (0.7304 \cdot sepal\_length) + (-0.6514 \cdot sepal\_width) + (1.4457 \cdot petal\_width)$$

Comparación de Métodos

Método	Intercepto	sepal_length	sepal_width	petal_width
Descenso Gradiente	-0.4126	0.7937	-0.6909	1.3707
Mínimos Cuadrados	-0.2527	0.7304	-0.6514	1.4457

Diferencia	0.1599	0.0633	0.0395	0.0750

Los coeficientes obtenidos por ambos métodos son muy similares, lo que confirma la robustez del modelo.
Las pequeñas diferencias se deben a que mínimos cuadrados es una solución exacta, mientras que descenso gradiente es aproximada, por lo que varían en su tasa de error.
El método de mínimos cuadrados proporciona estimaciones exactas en una sola operación, siendo más preciso para datasets de tamaño moderado.
El descenso del gradiente es más flexible y escalable, ideal para datasets muy grandes donde el método de mínimos cuadrados sería computacionalmente costoso.

F) Al utilizar regresión múltiple y agregar más variables predictoras mejoró significativamente la estimación en comparación con el modelo simple. El modelo de regresión múltiple explica el 96.8% de la variabilidad en la longitud del pétalo, mientras que el modelo simple explica solo el 92.7%. Esta diferencia del 4.45% refleja una mejora significativa en la precisión del modelo.

Las tres variables predictoras (sepal_length, sepal_width, y petal_width) resultaron ser estadísticamente significativas. El coeficiente de sepal_length muestra un efecto positivo (coeficiente = 0.7304), mientras que el de sepal_width es negativo (coeficiente = -0.6514). Petal_width sigue siendo el predictor más influyente, con el coeficiente más alto (1.4457).

En cuanto a la interpretación biológica, la inclusión de las dimensiones del sépalo aporta información adicional sobre el crecimiento de la flor, lo que complementa la relación ya observada entre la longitud y el ancho del pétalo. Esto indica que el crecimiento de los pétalos y los sépalos está relacionado y las dimensiones del sépalo también afectan la longitud del pétalo.

En conclusión, aunque la mejora en R^2 es modesta, el modelo múltiple permite una predicción más precisa y captura relaciones más complejas entre las variables, lo que lo hace más adecuado para obtener una estimación más detallada y biológicamente consistente.

Lautaro José Gubia
Facundo Nicolás Mendoza
Nicolás Agustín Muñoz