07 – Confounders

Autor: Peitsch, Pablo **mail:** preynosopeitsch@estudiantes.unsam.edu.ar
Github: @PPeitsch **Fecha: 2023-05-23**

En este caso se presenta una comparación de dos modelos de regresión lineal para la relación entre el ancho y largo del sépalo, para el ya conocido dataset Iris.

En la gráfica superior se puede observar un ajuste lineal a todos los datos, utilizando un modelo simple siguiendo una relación lineal entre el largo y ancho del sépalo (Sepal.Width = Sepal.Length). Si uno no tiene en cuenta las especies, en este caso marcadas en color, se puede apreciar cómo la recta atraviesa sobre la nube de puntos, lo cual podría indicar que, de alguna manera, el modelo ajusta. En cuanto a los estadísticos del modelo, que se pueden ver en la tabla a la izquierda del gráfico, el valor p para el intercept parece ser bueno. Sin embargo, el R-Squared arroja un valor muy lejano a 1, lo que indicaría que el modelo no explica la variabilidad de los datos. A su vez, el valor p para la variable Sepal.Width tampoco es bueno. Estos resultados podrían indicar que el modelo no representa muy bien a los datos.

Haciendo ajustes separados por especies, como se observa en la parte inferior, el resultado es notablemente mejor. El valor p para todas las variables es inferior a 4e-12 y el R-Squared se acerca mucho más a 1, lo que en ambos casos indica que las variables representan muy bien al modelo, mostrando una notable respecto al anterior.

En cuanto a los residuos, comparando ambos boxplot, en el segundo ajuste se puede observar una menor amplitud, o dicho de otra manera que la dispersión en los datos es más baja, indicando también una mejora en el modelo.

Por último, la comparación de los modelos utilizando ANOVA refuerza lo dicho anteriormente.

Comparando gráficamente los modelos se ve cómo la tendencia en ambos casos es inversa: para el primero el ancho del sépalo disminuye al aumentar el largo; en cambio, sucede lo opuesto en el segundo modelo. Se puede decir que si no se tiene en cuenta la variable Species, los datos podrían ajustarse erroneamente, lo cual indica que podría tratarse de una variable confundidora y, en este caso en particular, se puede apreciar cómo los estadísticos ayudaron a encontrar el modelo que mejor represente los datos.