Multicolinealidad en los modelos lineales generalizados, ¿el Chamucho de los estadísticos?

Dra. Eunice Campirán García.

August 2, 2023

Presentamos un conjunto de ejemplos interesantes para comprender las limitaciones de los métodos de detección de multicolinealidad en la regresión lineal múltiple y modelos lineales generalizados. En nuestros ejemplos, cada par de variables están altamente correlacionadas; no obstante, todas ellas deben incluirse en el modelo de regresión lineal, llegando a la conclusión de que la multicolinealidad por si misma no es necesariamente perjudicial en la estimación de los parámetros del modelo. Sorprendentemente, este problema se ha discutido en pocos trabajos; en cambio, la mayor parte de la literatura se dedica a la detección de la multicolinealidad y sus posibles soluciones, como eliminar un subconjunto de las variables altamente correlacionadas, opción que no es viable en los ejemplos presentados.

Para la construcción de nuestros ejemplos, utilizaremos la intuición detrás de la representación vectorial de las variables aleatorias y la correlación como el coseno de su ángulo, luego formalizamos nuestros resultados de manera algebraica. Finalmente, presentamos simulaciones donde los valores de R^2 y R^2_{adj} están cerca de uno cuando todas las variables altamente correlacionadas están presentes en nuestro modelo, mientras que estas estadísticas están cerca de cero cuando eliminamos una de ellas. Además, el valor p de la prueba F es cercano a cero en el modelo con todas las variables, y el valor p se distribuye uniformemente en el intervalo $[0, 1]$ cuando eliminamos una de ellas, lo que indica que la regresión es significativa solo en un 100α de las simulaciones, donde α es el nivel de significancia de la prueba F. Proporcionamos las funciones escritas en el software estadístico R para replicar nuestros resultados. Este trabajo muestra que necesitamos más herramientas que nos ayuden a distinguir entre multicolinealidad perjudicial e inofensiva en modelos lineales generalizados.

Palabras claves: **Multicolinealidad**; **Regresión Múltiple; Selección de Variables**; **Factor de Inflación de la Varianza**; **Número de Condición de una Matriz.**

Bibliografía.

Belsley, D. A. (1982). Assessing the presence of harmful collinearity and other forms of weak data through a test for signal-to-noise. Journal of Econometrics,

20(2), 211-253.

Chatterjee, S., & Hadi, A. S. (2013). Regression Analysis by Example. Wiley (5th ed.).

Gregorich, M., Strohmaier, S., Dunkler, D., & Heinze, G. (2021). Regression with Highly Correlated Predictors: Variable Omission Is Not the Solution. International Journal of Environmental Research and Public Health, 18(8), 4259.

O'Brien, R.M. (2017), Dropping Highly Collinear Variables from a Model: Why it Typically is Not a Good Idea*. Social Science Quarterly, 98: 360-375.