Tres patas de un gato

Lo de buscarle tres pies al gato, o tres patas, es un dicho muy popular. Parece que se dice que busca tres pies a un gato aquél que trata de demostrar alguna cosa imposible, generalmente con tretas y engaños. En realidad, el refrán inicial hacía referencia a buscar cinco pies en lugar de tres. Esto parece más lógico, ya que como los gatos tienen cuatro patas, encontrarles tres de ellas es cosa fácil, pero encontrar cinco es algo imposible, a no ser que consideremos la cola del gato como otro pie, lo cual no tiene mucho sentido.

Pero hoy no vamos a hablar de gatos con tres, cuatro o cinco pies. Vamos a hablar sobre algo un poco más etéreo, como son los modelos multivariables de regresión lineal múltiple. Este sí que es un gato con multitud de pies, pero nosotros nos vamos a fijar únicamente en tres de ellos que reciben los nombres de colinealidad, tolerancia y factor de inflación (o incremento) de la varianza. Que nadie se desanime, es más fácil de lo que puede parecer de entrada.

Ya vimos en una entrada anterior cómo los modelos de regresión lineal simple relacionaban dos variables entre sí, de forma que las variaciones de una de ellas (la variable independiente o predictora) podían servir para calcular cómo iba a variar la otra variable (la variable dependiente). Estos modelos se representaban según la ecuación y = a + bx, donde x es la variable independiente e y la dependiente.

Pues bien, la regresión lineal múltiple añade más variables independientes, de tal manera que permite hacer predicciones de la variable dependiente según los valores de las variables predictoras o independientes. La fórmula genérica sería la siguiente:

$y = a + bx_1 + cx_2 + dx_3 + \ldots + nx_n$, siendo n el número de variables independientes.

Una de las condiciones para que el modelo de regresión lineal múltiple funcione adecuadamente es que las variables independientes sean realmente independientes y no estén correlacionadas entre sí.

Imaginad un ejemplo absurdo en el que metemos en el modelo el peso en kilogramos y el peso en libras. Ambas variables variarán del mismo modo. De hecho el coeficiente de correlación, R, será 1, ya que prácticamente las dos representan la misma variable. Ejemplos tan tontos es difícil verlos en los trabajos científicos, pero hay otros menos evidentes (como incluir, por ejemplo la talla y el índice de masa corporal, que se calcula a partir del peso y de la talla) y otros que no son evidentes en absoluto para el investigador. Esto es lo que se llama colinealidad, que no es más que la

existencia de una asociación lineal entre el conjunto de las variables independientes.

La colinealidad es un grave problema para el modelo multivariable, ya que las estimaciones obtenidas por el mismo son muy inestables, al hacerse más difícil separar el efecto de cada variable predictora.

Pues bien, para determinar si nuestro modelo sufre de colinealidad podemos construir una matriz donde se muestran los coeficientes de correlación, R, de unas variables con otras. En aquellos casos en los que observemos R altos, podremos sospechar que existe colinealidad. Ahora bien, si queremos cuantificar esto recurriremos a las otras dos patas del gato que hemos comentado al inicio: tolerancia y factor de inflación de la varianza.

Si elevamos el coeficiente R al cuadrado obtenemos el coeficiente de determinación (R^2), que representa el porcentaje de la variación (o varianza) de una variable que es explicada por la variación en la otra variable. Así, nos encontramos con el concepto de tolerancia, que se calcula como el complementario de R^2 ($1-R^2$) y que representa la proporción de la variabilidad de dicha variable que no se explica por el resto de las variables independientes incluidas en el modelo de regresión.

De esta forma, cuanto más baja sea la tolerancia, más probable será que exista colinealidad. Suele considerarse que existe colinealidad cuando R^2 es superior a 0,9 y, por tanto, la tolerancia está por debajo de 0,1.

Ya solo nos queda la tercera pata, que es el factor de inflación de la varianza. Este se calcula como el inverso de la tolerancia (1/T) y representa la proporción de la variabilidad (o varianza) de la variable que es explicada por el resto de las variables predictoras del modelo. Como es lógico, cuanto mayor sea el factor de inflación de la varianza, mayor será la probabilidad de que exista colinealidad. Generalmente se considera que existe colinealidad cuando el factor de inflación entre dos variables es mayor de 10 o cuando la media de todos los factores de inflación de todas las variables independientes es muy superior a uno.

Y aquí vamos a dejar los modelos multivariables por hoy. Ni que decir tiene que todo lo que hemos contado en la práctica se hace recurriendo a programas informáticos que nos calculan estos parámetros de manera sencilla.

Hemos visto aquí algunos de los aspectos de la regresión lineal múltiple, quizás el más utilizado de los modelos multivariables. Pero hay otros, como el análisis multivariante de la varianza (MANOVA), el análisis factorial o el análisis por conglomerados o clústeres. Pero esa es otra historia…