Regresión Múltiple

Miguel Ernesto Campos Ramos

Modelo

Un modelo relaciona una o varias variables que hay que explicar Y a unas variables explicativas X, por una relación funcional Y = F (X)

Un modelo físico es un modelo explicativo sostenido por una teoría.

Un modelo estadístico, al contrario, es un modelo empírico nacido de datos disponibles, sin conocimientos a priori sobre los mecanismos en juego. Podemos sin embargo integrar en eso ecuaciones físicas (en el momento del pretratamiento de datos).

Disponemos de n de observaciones (i = 1…, n) de p variables. La ecuación de regresión se escribe:

donde

- ε_i *es el error del modelo;*
- $a_0, a_1, \ldots, a_p$ *son los coeficientes del modelo que hay que estimar.*

El cálculo de los coeficientes a_j *y del error del modelo, a partir de las observaciones, es un problema bien dominado*

Más delicado es la elección de las variables que entran en este modelo. Puede ser postulado o no postulado.

Modelo postulado

Solo los coeficientes del modelo precedente de regresión son dirigidos por los datos, la estructura polinómica del modelo es impuesta por el utilizador (según su peritaje del problema), que postula a priori:

- *El tipo de modelo: lineal o polinómico, y el grado del polinomio,*
- *las variables que entrarán en el modelo.*

El problema de la selección de las variables explicativas

Cuando el número de variables explicativas es grande, puede hacerse que ciertas variables sean correlacionadas. En este caso hay que eliminar los doblones. El software utiliza para hacerlo métodos de selección paso a paso (ascendientes, descendentes o mixtos).

Sin embargo, la calidad del modelo final repone en gran parte en la elección de las variables, y del grado del polinomio.

Modelo no postulado

La selección se produce antes del cálculo de los coeficientes de la regresión según el principio siguiente:

Buscamos el factor o la interacción o la función mejor correlada a la respuesta. Habiéndolo encontrado, buscamos el factor o la interacción mejor correlada al residuo no explicado por la correlación precedente; etc. Este método pretende no contar dos veces la misma influencia, cuando los factores son correlados, y a ordenarlos por importancia decreciente.

La lista por orden de importancia decreciente encontrada y clasificada, no puede contar más términos que desconocidas (n). Si se guarda sólo un término en el modelo, deberá ser la primera de la lista. Si se guarda dos, serán ambos primeros, etc.

En efecto ya que cada uno de los términos de la lista explica el residuo no explicado por los precedentes, los últimos explican posiblemente solo el ruido. ¿ Cuál criterio de parada escoger?

El número de términos conservados en el modelo puede ser, por ejemplo, el que minimiza el error estándar de predicción SEP (Standard error of Prediction), o el que maximiza el F de Fisher. Este número de término puede también ser escogido por el utilizador a partir de consideraciones físicas.

Descomposición armónica

Un modelo no postulado será también eficaz en la descomposición armónica de las series.

En efecto, el principio se aplica también bien en caso de muestreo irregular (donde los métodos de tipo media móvil, ARIMA o Box y Jenkins son hechos caer en falta) que en los casos no estacionarios (donde Análisis armónico no se aplica). Permite descubrir y desenredar las interferencias de ciclos diversos y estacionalidad con roturas de tendencias en escalón, en V, roturas logísticas, motivos periódicos, y acontecimientos accidentales tales como picos aislados o pedazos de ondas.