

Práctica 1: Métodos de selección de variables

Modelos Lineales

Grados en Estadística y Matemáticas

Selección de variables

En el modelo que hemos ajustado no todas las variables predictoras, $X_1, \dots, X_p$, tienen por qué ser significativas a la hora de explicar el comportamiento de la variable respuesta Y .

Podemos reducir nuestro modelo eliminando aquellas variables que no resulten significativas o que no aporten información al modelo. Con ello podemos dar también solución al problema de la multicolinealidad.

Métodos de selección por pasos

- **Pasos hacia adelante o forward.** Partimos del modelo sin variables y vamos introduciendo la que sea más significativa (menor p-valor).
- **Hacia atrás o backward.** Partimos de un modelo con todas las variables y vamos eliminando la menos significativa (mayor p-valor).

Métodos de selección globales

- **Basado en R_A^2** : De todos los modelos que podemos construir, considerando un subconjunto de $X_1, \dots, X_p$ como variables predictoras, nos quedamos con el modelo con mayor R_A^2 .
- **Basado en el Criterio de Información de Akaike (AIC) o en el Criterio de Información Bayesiano (BIC)**: Ambos valores se calculan a partir de $Lmax = L(\hat{\beta}, \tilde{\sigma}^2)$, el valor máximo de la función de verosimilitud de dicho modelo. Se definen como

$$AIC = 2k - 2 \log(Lmax) \quad , \quad BIC = k \log(n) - 2 \log(Lmax)$$

siendo k el número de parámetros del modelo, incluyendo σ^2 .

De todos los modelos que podemos construir, considerando un subconjunto de $X_1, \dots, X_p$ como variables predictoras, nos quedamos con el modelo con menor AIC o BIC.

Métodos de selección globales

- **Basado en el Coeficiente C_p de Mallows:** Para un modelo concreto, se calcula mediante la comparación entre los estimadores de σ^2 en dicho modelo y en el modelo con todas las variables regresoras. Se define como

$$C_p = q + (n - q) \left(\frac{\hat{\sigma}_q^2}{\hat{\sigma}^2} - 1 \right)$$

siendo $q \leq p$ el número de variables regresoras en el modelo y $\hat{\sigma}_q^2$ el estimador de σ^2 en el modelo. De todos los modelos posibles nos quedaremos con el que presente un menor coeficiente C_p .

Ejemplo 1

Se realiza un experimento para determinar el calor desarrollado en la fabricación de cemento en función de los porcentajes de 4 compuestos activos que se utilizan en el proceso.

Para ello se elige una muestra de 13 cementos, midiéndose el calor desarrollado de cals/g (y) y los porcentajes de los compuestos referidos (x_1, x_2, x_3, x_4).

Los datos de la muestra se proporcionan en el archivo `cemento.dat`.