Introducción a la Estadística y Ciencia de Datos - 2023

Práctica 7 - Métodos de regularización

---

1. Simular los datos con tamaño de muestra $n = 100$,
   $X_1 \sim U(0;5)$
   $X_2 \sim U(0;5)$
   $X_3 = 2X_1 + U_3$, con $U_1 \sim U(-0{,}1; 0{,}1)$
   $X_4 = -X_1 + U_4$, con $U_4 \sim U(-0{,}1; 0{,}1)$
   $Y = 8X_1 - 5X_2 + X_3 + 4X_4 + 5 + \varepsilon$, con $\varepsilon \sim N(0,1)$
   $X_1, X_2, U_3, U_4, \varepsilon$ son variables independientes
   Es decir, la respuesta Y depende de dos covariables independientes, X_1, X_2 y las otras dos X_3, X_4 son casi colineales con X_1.

   a) Repetir el experimento $B = 1000$ veces. Cada vez, ajustar un modelo lineal por mínimos cuadrados y otro modelo usando la regularización Ridge, calcular el desvío estándar de los estimadores y la proporción de veces que cada coeficiente resultó significativo. Comparar resultados para los dos modelos.

   b) Repetir el ítem anterior cambiando Ridge por Lasso.

2. Los datos en el archivo `peak.txt` son datos simulados que corresponden al caudal de agua (Y) en distintas cuencas de ríos después de episodios de tormenta. Las covariables son:
   X_1= área de la cuenca
   X_2= área impermeable al agua
   X_3= pendiente promedio del terreno
   X_4= máxima longitud de los afluentes de la cuenca
   X_5= índice de absorción del agua (0= absorción completa, 100= no absorción)
   X_6= capacidad de depósito del suelo
   X_7= velocidad de infiltración del agua en el suelo
   X_8= cantidad de lluvia caída
   X_9= tiempo durante el cual la cantidad de lluvia excedió 0.25 pulgada por hora

   Observación:

   El R^2_{ajustado} es una versión corregida del coeficiente de correlación múltiple R^2 que penaliza a un modelo lineal de acuerdo a la cantidad de coeficientes lineales que utiliza, esencialmente dividiendo a la suma de cuadrados del numerador y la del denominador por sus grados de libertad. En este sentido, constituye una manera de comparar modelos con distinto número de parámetros, eligiendo aquel con mayor R^2_{ajustado}. Más precisamente, supongamos que para el vector de respuestas $\mathbf{Y}$, un modelo lineal postula $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\theta}$, donde $\mathbf{X} \in \mathbb{R}^{n\times k}$, $\boldsymbol{\theta} \in \mathbb{R}^k$ y el vector de predichos es $\widehat{\mathbf{Y}}$. Luego, si $\overline{\mathbf{Y}}$ es el vector cuyas componentes son iguales a $\overline{Y}$, entonces se define

$$R^2_{\text{ajustado}} = 1 - \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2/(n-k-1)}{\|\mathbf{Y} - \overline{\mathbf{Y}}\|^2/(n-1)} = 1 - (1-R^2)\frac{n-1}{n-k-1}$$

   Cabe destacar que el R^2_{ajustado} es reportado automáticamente en el summary del `lm`.

a) Definir $\ln(Y)$ como variable dependiente y los logaritmos naturales de las covariables y una intercept para realizar un ajuste lineal. Calcular el estimador de mínimos cuadrados de los parámetros y para cada uno de ellos testear la hipótesis que sea 0. ¿ Cuáles son significativamente distintos de 0? ¿Cuáles son las variables que eliminaría en primera instancia para simplificar el modelo? ¿Es la regresión significativa?

b) Mediante el procedimiento Stepwise, seleccionar el mejor modelo:

i) usando como medida de bondad de ajuste el R^2_{ajustado}.

ii) calculando el error cuadrático medio de validación cruzada.

c) Ajustar nuevamente el modelo propuesto en el ítem $2a$, pero ahora usando penalización Lasso. Utilizando el lambda de 1 desvío estándar, observar cuáles son las variables eliminadas por el procedimiento, ¿coinciden con las variables eliminadas en el ítem anterior? Una vez realizado el ajuste, calcular el ECM de validación cruzada y decidir, de todos los modelos propuestos en el ejercicio, el modelo óptimo para predecir a Y.