Curso: Cómputo Estadístico

Tarea

Instrucciones

1. Los ejercicios se entregarán con el script de **R** que generen y con un reporte que incluya los resultados obtenidos y las respuestas solicitadas. El script y el reporte se subirán a la plataforma.
2. La fecha límite de entrega de la tarea es el **domingo 2 de noviembre a las 23:00**

Ejercicios

1. Generación de datos simulados y aplicación de los métodos de selección de subconjuntos
   (a) Usa una función en **R** para generar una variable predictora $\boldsymbol{X}$ de longitud $n = 100$, así como un vector de ruido $\boldsymbol{\epsilon}$ de tamaños $n = 100$
   (b) Genera un vector de respuesta $\mathbf{Y}$ de longitud $n = 100$ de acuerdo al modelo
   $$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$
   donde $\beta_0, \beta_1, \beta_2$ y β_3 son constantes de tu elección
   (c) Utiliza la funcion **regsubsets ()** para realizar la seleccion de los mejores subconjuntos con el fin de elegir el mejor modelo que contenga los predictores $X, X^2, X^3, \ldots, X^{10}$, ¿ Cuál es el mejor modelo obtenido según el **AIC**, **BIC** y el $\boldsymbol{R^2}$ ajustado? Muestra algunas gráficas que proporcionen evidencia de tu respuesta y reporta los coeficientes del mejor modelo obtenido
   (d) Repite (c) usando la selección forward stepwise y backward stepwise,¿ Cómo se compara tu respuesta con los resultados obtenidos en (c)?
2. Se ha visto que a medida que aumenta el número de características de un modelo, el error de entrenamiento disminuirá necesariamente, pero el error de prueba no. Explorar esto con datos simulados
   (a) Genera un conjunto de datos con $p = 20$ características, $n = 1000$ observaciones y un vector de respuesta cuantitativo generado de acuerdo con el modelo
   $$Y = X\beta + \epsilon$$
   donde β tiene algunos elementos que son exactamente iguales a cero.
   (b) Divide tu conjunto de datos en un conjunto de entrenamiento que contenga 100 observaciones y un conjunto de pruebas que contenga 900 observaciones.
   (c) Realiza *la selección del mejor subconjunto* sobre el conjunto de entrenamiento y grafica el error de entrenamiento MSE asociado con el mejor modelo en cada tamaño.
   (d) Grafica el error de prueba MSE asociado con el mejor modelo de cada tamaño

(e) ¿Para qué tamaño de modelo el error de prueba MSE toma su valor mínimo? Comenta tus resultados. Si toma su valor mínimo en un modelo que sólo contiene una interceptación o un modelo que contenga todas las características, entonces juega con la forma en la que estás generando los datos en (a) hasta que aparezca un escenario en el que el error de prueba MSE se minimiza para un tamaño de modelo intermedio.

(f) ¿Cómo se compara el modelo con el que se minimiza el error de prueba con el modelo verdadero utilizado para generar los datos? Comenta sobre los valores de los coeficientes