Examen final de Regresión Avanzada - 07/12/2024

La base de datos con la que trabajaremos está en el archivo **salarios.csv.** Las variables que utilizaremos de esta base son las siguientes:

salario: salario promedio por hora

educ: años de educación

exper: años de experiencia potencial

ternure: años con el empleador actual (antigüedad)

nonwhite: es igual a 1 si la persona no es blanca

female: es igual a 1 si la persona es mujer

married: es igual a 1 si la persona es casada

numpdep: número de dependientes

smsa: 1 si vive en un área metropolitana

south: 1 si vive en el sur de USA

west: 1 si vive en el este de USA

construc: 1 si trabaja en la construcción

clerocc: 1 si es del clero

servocc: 1 si tiene ocupación en servicios

profocc: 1 si tiene un trabajo profesional

lwage: logaritmo del salario, log(wage)

expersq: raíz cuadrada de los años de experiencia potencial,

exper2: cuadrado de la experiencia potencial

ternure2: el cuadrado de la antigüedad en el trabajo actual

ternuresq: raíz cuadrada de la antigüedad en el trabajo actual

Nuestro interés es utilizar estas variables para explicar el salario.

a) Construya el mejor modelo lineal simple y realice el análisis diagnóstico del mismo. En caso de ser necesario transforme la variable respuesta. Concluya.

b) Mediante selección de variables elija el mejor modelo multivariado para explicar salario. Estudie la presencia de multicolinealidad. Si la hubiera aplique alguna metodología para evitarla.

c) Analice la pertinencia de un modelo gamlss para explicar Salario.

d) Divida los salarios en dos grupos:

   Grupo 1: menores al cuartil 3

   Grupo 2: iguales o mayores al cuartil 3

   Construya un modelo logístico e interprete los coeficientes. Analice la bondad de ajuste del mismo. Concluya.

e) Compare este método con otro método de clasificación y concluya.