UNIVERSIDAD NACIONAL DE COLOMBIA

FACULTAD DE CIENCIAS, SEDE BOGOTÁ
DEPARTAMENTO DE ESTADÍSTICA
ANÁLISIS DE REGRESIÓN

Ejercicio en clase 5. Selección de modelos y enfoque predictivo.

1. Considere la base de datos **Hitters** incluida en el paquete **ISLR**. Esta base contiene la información de las estadísticas de desempeño de 322 bateadores de las ligas mayores de Estados Unidos entre 1986 y 1987, al igual que los salarios que devengaron en 1987.

    a) Explore la base de datos y las variables que contiene. Identifique los individuos para los cuales no se tiene información del salario y remuévalos (ya que la idea será predecir el salario en función de los otros factores).
    b) Evalúe todas las posibles regresiones y analice qué variables son incluidas para cada modelo en cada tamaño.
    c) Compare los modelos de diferentes tamaños usando criterios como AIC, BIC, R2, etcétera; y defina qué tamaño es óptimo en cada caso.
    d) Haga la regresión forward y backward y compare con los resultados anteriores.

2. Usando la misma base de datos, vamos a seleccionar el mejor modelo en términos de su habilidad predictiva.
    a) Separe el conjunto de datos al azar en un 70% de entrenamiento y 30% de testeo.
    b) Encuentre el mejor modelo de en términos de su habilidad predictiva. Use el error cuadrático medio de predicción para tal fin.
    c) Reestime el modelo con todos los datos.

3. Repita el ejercicio 2, pero esta vez use validación cruzada en 10 etapas.

4. Retomando el contexto del problema 2, construya un modelo de regresión ridge y uno LASSO y determine cuál tiene mejor habilidad predictiva. Use las mismas condiciones y criterios del problema 2. Además, encuentre lambda mediante validación cruzada dentro de la etapa de entrenamiento.