Індивідуальне завдання: вибір лінійної моделі та регуляризація.

Встановіть значення змінної variant: сума номера групи помноженого на 25 і порядкового номеру студента в списку групи (групі ПМОм-11 відповідає номер 0, групі ПМІм-11 відповідає номер 1, групі ПМІм-12 відповідає номер 2, групі ПМІм-13 відповідає номер 3). Далі встановіть set.seed(variant) та згенеруйте значення змінної redundant як заокруглене до цілого (для заокруглення можна використати функції floor або round) випадкове число з рівномірного на інтервалі (номер групи + 5, 25 – номер групи) розподілу (функція runif).

1. Модифікуйте дані Auto наступним чином: встановивши seed, що дорівнює значенню змінної variant, видаліть redundant % спостережень з допомогою функції sample. Розбийте набір даних на навчальний та тестовий набори попередньо встановивши seed, що дорівнює значенню змінної variant. До тестового набору включіть 2*redundant % усіх спостережень. Передбачимо значення змінної mpg на основі всіх інших змінних. Використайте на тренувальному наборі даних: лінійну модель на основі методу найменших квадратів, модель гребеневої регресії та модель ласо, вибравши λ шляхом перехресної перевірки, моделі PCR та PLS, вибравши M шляхом перехресної перевірки. Для кожної з моделей оцініть тестову помилку. Прокоментуйте отримані результати.

2. Встановивши попередньо seed, що дорівнює значенню змінної variant, використайте функцію rnorm() та згенеруйте предиктор X довжиною n = 100 * (1+variant %/% 10) з середнім μ = [variant/5] + 1 та середньоквадратичним відхиленням σ = [(2*variant)^(1/2)]+1), та вектор залишків ε такої ж довжини n з параметрами μ = 0 та σ = 1). Виберіть β_0, β_1, β_2 і β_3 (попередньо встановивши seed, що дорівнює значенню змінної variant) як реалізації рівномірно розподіленої випадкової величини на відрізку [-10, 10] заокруглені до найближчого цілого та обчисліть

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon.$$

Використовуючи функцію regsubsets() виберіть найкращу модель методом вибору найкращої підмножини з множини предикторів X, X^2, . . ., X^{10}. Яка модель найкраща за показниками C_p, BIC і скорегований R^2? Використайте методи покрокового вибору вперед та назад та порівняйте результати з результатами вибору найкращої підмножини. Пристосуйте модель ласо до згенерованих даних, використовуючи X, X^2, . . ., X^{10} як предиктори. Використайте перехресну перевірку для вибору значення λ. Наведіть отримані оцінки коефіцієнтів моделі та обгрунтуйте отримані результати.