Індивідуальне завдання: методи ресемплінгу.

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. Розглянемо набір даних Boston з бібліотеки MASS. Модифікуйте ці дані наступним чином: встановивши seed, що дорівнює значенню змінної variant, видаліть redundant % спостережень з допомогою функції sample. Використовуючи модифіковані дані, пристосуйте модель логістичної регресії для передбачення у вибраному районі рівня злочинності більшого чи меншого за середній на основі змінних nox та rad. Оцініть тестову помилку цієї моделі логістичної регресії, використовуючи метод валідаційного набору (використати розбиття 50 на 50, попередньо скинувши seed). Повторіть попередню процедуру три рази, використовуючи нові розбиття вибірки на навчальний та тестовий набори. Прокоментуйте отримані результати. Розгляньте модель логістичної регресії для передбачення у вибраному районі рівня злочинності більшого чи меншого за середній на основі змінних nox, rad та medv. Оцініть тестову помилку для цієї моделі, використовуючи метод валідаційного набору описаний вище. Прокоментуйте, чи призводить включення нової змінної до зменшення тестової помилки.

2. Модифікуйте дані Auto наступним чином: встановивши seed, що дорівнює значенню змінної variant, видаліть redundant % спостережень з допомогою функції sample. На основі цього набору даних обчисліть оцінку середнього змінної mpg. Оцініть стандартну похибку цієї оцінки. Тепер оцініть стандартну похибку розглянутої вище оцінки середнього за допомогою бутстрапу та порівняйте з попередньо отриманим результатом. Обчисліть оцінку для медіани та десятого процентиля змінної mpg. Оцініть стандартні помилки отриманих оцінок допомогою бутстрапу.

3. Встановіть seed, що дорівнює значенню змінної variant та створіть змодельований набір даних наступним чином:
> x = rnorm (100)
> y = variant*x – ((redundant*40)/variant) * x ^ 2 + rnorm (100)
Встановіть seed, що дорівнює значенню змінної variant та обчисліть оцінки тестових помилок методом LOOCV, для наступних чотирьох моделей

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

$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$
$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \varepsilon$

Яка з моделей має найменшу тестову помилку LOOCV? Чи це відповідає очікуванням? Поясніть свою відповідь.