1. Ми використовували логістичну регресію для прогнозування ймовірності дефолту на основі income та balance з даних Default. Оцініть тестову помилку цієї моделі логістичної регресії, використовуючи метод валідаційного набору.

1.1. Побудуйте логістичну регресійну модель, яка використовує income та balance для передбачення default.

1.2. Використовуючи метод валідаційного набору, оцініть тестову помилку цієї моделі. Для цього виконайте такі дії:

1.2.1. Розділіть вибірку на навчальний та тестовий набори.

1.2.2. Оцініть логістичну регресійну модель, використовуючи навчальну вибірку.

1.2.3. Спрогнозуйте дефолт-статус для кожної людини в тестовій вибірці на основі передбачення апостеріорної ймовірності дефолту для цієї людини та встановлення статусу дефолт, якщо отримана ймовірність перевищує значення 0,5.

1.2.4. Оцініть тестову помилку на валідаційній множині шляхом обчислення частки статусу осіб, які неправильно класифіковані.

1.3. Повторіть 1.2 три рази, використовуючи три різні розбиття вибірки на навчальний та тестовий набори. Прокоментуйте отримані результи.

1.4. Розгляньте модель логістичної регресії, яка передбачає ймовірність дефолту за допомогою змінних income, balance та фіктивної змінної для student. Оцініть тестову помилку для цієї моделі, використовуючи використаний вище підхід. Прокоментуйте, чи призводить включення фіктивної змінної для student до зменшення тестової помилки.

2. Продовжимо дослідження логістичної регресії для прогнозування ймовірності дефолту на основі income та balance з даних Default. Оцініть стандартне відхилення оцінок параметрів цієї моделі логістичної регресії, використовуючи: бутстрап; стандартну формулу функції glm().

2.1. Використовуючи функції summary() та glm(), визначте оцінку середньоквадратиного відхилення параметрів логістичної регресії, яка використовує income та balance для оцінки ймовірності дефолту.

2.2. Напишіть функцію boot.fn (), яка приймає на вхід набір даних та індекси спостережень для використання і виводить оцінки коефіцієнтів логістичної регресії, яка використовує income та balance для оцінки ймовірності дефолту.

2.3. Використайте функцію boot() разом із функцією boot.fn() для оцінки середньоквадратиного відхилення параметрів логістичної регресії, яка використовує income та balance для оцінки ймовірності дефолту.

2.4. Прокоментуйте отримані результати.

3. Обчисліть оцінку тестової помилки методом LOOCV, використовуючи лише функції glm (), predict.glm () та цикл for.

Застосуйте такий підхід для того, щоб обчислити оцінку тестової помилки методом LOOCV для логістичної регресійної моделі на наборі даних Weekly.

3.1. Побудуйте модель логістичної регресії, яка передбачає Direction за допомогою змінних Lag1 та Lag2.

3.2. Побудуйте модель логістичної регресії, яка передбачає Direction за допомогою змінних Lag1 та Lag2, використовуючи всі спостереження, крім першого.

3.3. Використайте модель з 3.2, щоб передбачити Direction для першого спостереження. Це можна зробити, передбачивши, що перше спостереження піде вгору, якщо P(Direction = "Up" | Lag1, Lag2)> 0,5. Чи правильно класифіковано це спостереження?

3.4. Напишіть цикл for від i = 1 до i = n, де n - число спостережень в наборі даних, який виконує наступні кроки:

Будує модель логістичної регресії, яка передбачає Direction за допомогою змінних Lag1 та Lag2, використовуючи всі спостереження, крім і-ого.

Обчислює апостеріорну ймовірність руху ринку вгору для i-го спостереження.

Використовує апостеріорну ймовірність для i-го спостереження для прогнозування, чи рухатиметься ринок вгору чи ні.

Визначає, чи допущена помилка при прогнозуванні Direction для i-го спостереження. Якщо була допущена помилка, то вказуємо це як 1, а в іншому випадку - як 0.

3.5. Обчисліть середнє з n чисел, отриманих у 3.4.4, для того, щоб отримати оцінку LOOCV для тестової помилки. Прокоментуйте результати.

4. Проведіть перехресну перевірку на змодельованому наборі даних.

4.1. Створіть змодельований набір даних наступним чином:

> set .seed (1)

> y = rnorm (100)

> x = rnorm (100)

> y = x-2 \* x ^ 2 + rnorm (100)

Що в цьому наборі даних *n* і що *p*? Випишіть модель, що використовувалася для генерації даних у формі рівняння.

4.2. Побудуйте діаграму розсіювання X vs Y. Прокоментуйте результати.

4.3. Встановіть random.seed та обчисліть оцінки тестових помилок методом LOOCV, для наступних чотирьох моделей

4.3.1. *Y* = *β*0 + *β*1*X* + *ε*

4.3.2. *Y* = *β*0 + *β*1*X* + *β*2*X*2 + *ε*

4.3.3. *Y* = *β*0 + *β*1*X* + *β*2*X*2 + *β*3*X*3 + *ε*

4.3.4. *Y* = *β*0 + *β*1*X* + *β*2*X*2 + *β*3*X*3 + *β*4*X*4 + *ε*

4.4. Повторіть 4.3, використовуючи інший random.seed, і повідомте про свої результати. Ваші результати такі самі, як і в 4.3? Чому?

4.5. Яка з моделей у 4.3. мала найменшу тестову помилку LOOCV? Чи це відповідає очікуванням? Поясніть свою відповідь.

4.6 Прокоментуйте статистичну значимість оцінок коефіцієнтів кожної з моделей розглянутих у 4.3. Чи узгоджуються ці результати із зробленими висновками на основі результатів перехресної перевірки?

5. Розглянемо набір даних Boston з бібліотеки MASS.

5.1. На основі цього набору даних обчисліть оцінку середнього змінної medv.

5.2. Оцініть стандартну похибку цієї оцінки. Прокоментуйте результат.

5.3. Тепер оцініть стандартну похибку розглянутої вище оцінки середнього за допомогою бутстрапу. Порівняйте її з результатом 5.2?

5.4. На основі бутстрап оцінки побудуйте 95% довіри для середнього значення змінної medv. Порівняйте це з результатами отримано за допомогою t.test(Boston$medv).

5.5. На основі цього набору даних обчисліть оцінку для медіани змінної medv.

5.6. Для оцінки стандартної похибки оцінки медіани, на жаль, немає простої формули. Тому оцініть стандартну помилку оцінки медіани змінної medv за допомогою бутстрапу. Прокоментуйте свої результати.

5.7. На основі цього набору даних обчисліть оцінку десятого процентиля змінної medv (використайте функцію quantile ()).

5.8. Використайте бутстрап, щоб оцінити стандартну похибку десятого процентиля змінної medv. Прокоментуйте ваші результати.