1. На основі згенерованих даних застосуємо вибір найкращої підмножини.

1.1. Використовуючи функцію rnorm() згенеруйте предиктор *X* довжиною *n* = 100, та вектор залишків *ε* такої ж довжини *n* = 100.

1.2. Згенеруйте вектор залежних змінних *Y* довжини *n* = 100 відповідно до моделі

*Y* = *β*0 + *β*1*X* + *β*2*X*2 + *β*3*X*3 + *ε*,

де *β*0, *β*1, *β*2 і *β*3 константи на ваш вибір.

1.3. Використовуючи функцію regsubsets() виберіть найкращу модель методом вибору найкращої підмножини з множини предикторів *X*, *X*2, . . ., *X*10. Яка модель найкраща за показниками *Cp*, *BIC* і скорегований *R*2? Наведіть декілька графіків на підтвердження своєї відповіді та вкажіть оцінки коефіцієнтів найкращої моделі.

1.4. Повторіть 1.3, використовуючи методи покрокового вибору вперед та назад. Порівняйте отримані результати з 1.3.

1.5. Пристосуйте ласо модель до згенерованих даних, використовуючи *X*, *X*2, . . ., *X*10 як предиктори . Використайте перехресну перевірку для вибору значення λ. Побудуйте графіки помилки перехресної перевірки як функції від λ. Наведіть отримані оцінки коефіцієнтів моделі та обгрунтуйте отримані результати.

1.6. Згенеруйте вектор залежних змінних *Y* відповідно до моделі

*Y* = *β*0 + *β*7*X*7 + *ε*,

і застосуйте метод найкращого вибору підмножини і ласо. Обгрунтуйте отримані результати.

2. На основі даних College передбачимо кількість отриманих заяв.

2.1. Розбийте набір даних на навчальний та тестовий набори.

2.2. Оцініть лінійну модель, використовуючи метод найменших квадратів на навчальному наборі, та обчисліть тестову помилку.

2.3. Пристосуйте модель гребеневої регресії до тренувального набору, вибравши λ шляхом перехресної перевірки. Обчисліть тестову помилку.

2.4. Пристосуйте модель ласо до тренувального набору, вибравши λ шляхом перехресної перевірки. Обчисліть тестову помилку. Яка кількість ненульових оцінок коефіцієнтів.

2.5 Пристосуйте модель PCR до тренувального набору, причому М виберіть шляхом перехресної перевірки. Яке отримане значення М? Обчисліть отриману помилку тесту.

2.6. Пристосуйте модель PLS до тренувального набору, причому М вибиріть шляхом перехресної перевірки. Яке отримане значення М? Обчисліть отриману помилку тесту.

2.7. Прокоментуйте отримані результати. Наскільки точно ми можемо передбачити кількість отриманих заявок на коледж? Чи велика різниця між тетсовими помилками, що виникають внаслідок розглянутих п’яти підходів?

3. Ми бачили, що зі збільшенням кількості предикторів, що використовуються в моделі, навчальна помилка обов'язково зменшиться, але тестова - не обов'язково. Дослідимо це на згенерованих даних.

3.1. Сформуйте набір даних з *p* = 20 ознаками, *n* = 1000 спостереженнями, і пов'язаний з ним вектор залежних змінних відповідно до моделі

*Y* = *Xβ* + *ε*,

де вектор *β* має деякі елементи, які точно дорівнюють нулю.

3.2. Розділіть свій набір даних на навчальний набір, що містить 100 спостережень та тестовий набір, що містить 900 спостережень.

3.3. Використайте метод вибору найкращої підмножини на навчальному наборі та побудуйте графік навчального MSE, який відповідає найкращій моделі кожного розміру.

3.4. Побудуйте графік тестового MSE, , який відповідає найкращій моделі кожного розміру.

3.5. Для якого розміру моделі тестовий MSE приймає мінімальне значення? Прокоментуйте отримані результати. Якщо ця величина набуває свого мінімального значення для нульової моделі, чи моделі, що містить усі предиктори, поверніться до 3.1. та поекспериментуйте з генерацією поки не отримаєте сценарій за якого тестовий MSE досягає мінімуму для якоїсь проміжної моделі.

3.6. Як співвідносяться модель, що мінімізує тестовий MSE та справжня модель, яка використовувалася для генерації даних? Прокоментуйте значення оцінок коефіцієнтів.

3.7. Побудуйте графік для відображення величини



для всіх значень *r*, де  - оцінка *j*-ого коефіцієнта для найкращої моделі, що містить *r* коефіцієнтів. Прокоментуйте результати. Порівняйте отриманий графік з графіком тестового MSE з 3.4?

4. Спробуємо передбачити рівень злочинності на основі набору даних Boston.

4.1. Застосуйте методи вибору моделі регресії, розглянуті раніше, такі як вибір найкращої підмножини, ласо, гребенева регресія та PCR. Представте та обговоріть результати щодо підходів, які ви використовуєте.

4.2. Запропонуйте модель, яка мала б добре працювати і обґрунтуйте свою відповідь. Для оцінки якості моделі використайте помилки валідаційної множини, перехресної перевірки, або іншу розумну альтернатива (не використовуйте навчальну помилку).

4.3. Чи включає обрана модель всі предиктори? Чому?