1. Додатково проаналізуєте набір даних Wage.

1.1. Використайте поліноміальну регресію для прогнозування wage за age. Використайте перехресну перевірку для вибору оптимального степеня d для полінома. Який степінь було обрано, і як це співвідноситься з результатами перевірки гіпотез з використанням ANOVA? Побудуйте графік отриманого поліному пристосованого до даних.

1.2. Використайте східчасту функцію для прогнозування wage за age та проведіть перехресну перевірку для вибору оптимальної кількості розрізів. Побудуйте графік з отриманими результатами.

2. Набір даних Wage містить інші змінні такі як, сімейний стан (maritl), робочий клас (jobclass) тощо. Дослідіть зв’язки між деякими з цих інших предикторів та wage, а також використовуючи нелінійні методи пристосуйте гнучкі моделі до даних. Побудуйте графіки отриманих результатів, та підсумуйте свої висновки.

3. Пристосуйте деякі нелінійні моделі на наборі даних Auto. Чи є якісь докази нелінійних взаємозв’язків в цьому наборі даних? Побудуйте кілька інформативних графіків, щоб обґрунтувати свою відповідь.

4. Використаємо змінні dis (зважене середнє відстані до п'яти центрів зайнятості в Бостоні) та nox (концентрація оксидів азоту у частинах на 10 мільйонів) з набору даних Boston. Розглянемо dis як предиктор та nox як залежну змінну.

4.1. Використовуючи функцію poly(), встановіть кубічну поліноміальну регресію для передбачення nox за допомогою dis. Опишіть результати регресії та побудуйте графік даних та поліноміальної регресії.

4.2. Побудуйте поліноміальні моделі для різних степенів (скажімо, від 1 до 10), і наведіть їхні RSS.

4.3. Використайте перехресну перевірку або інший підхід для вибору оптимального степеня для поліноміальної регресії та поясніть отримані результати.

4.4. Використовуючи функцію bs (), пристосуйте сплайн регресію для прогнозування nox за допомогою dis. Опишіть результати отримані з використанням чотирьох ступенів свободи. Як ви вибрали вузли? Побудуйте графік отриманої моделі.

4.5. Пристосуйте сплайн регресію для діапазону ступенів свободи, і побудуйте графік результатів. Наведіть відповідні RSS. Опишіть отримані результати.

4.6. Використайте перехресну перевірку або інший підхід, щоб вибрати найкращий ступінь свободи для сплайн регресії на цих даних. Опишіть свої результати.

5. Використайте набір даних College.

5.1. Розбийте дані на навчальний та тестовий набори. Використайте Outstate як залежну змінну, а інші змінні як предиктори. Виконайте покроковий вибір вперед на навчальному наборі, щоб визначити задовільну модель, яка використовує лише підмножину предикторів.

5.2. Оцініть УАМ модель на навчальних даних, використовуючи Outstate як залежну змінну та ознаки обрані на попередньому кроці як предиктори. Побудуйте графік результатів та поясніть свої висновки.

5.3. Застосуйте модель на тестовому наборі даних, поясніть отримані результати.

5.4. Для яких змінних, якщо такі є, є докази нелінійності взаємозв'язку з залежною змінною?

6. УАМ, як правило, оцінюють на основі методу підгонки. Розглянемо цей метод на основі множинної лінійної регресії. Припустимо, що ми хочемо використати множинну лінійну регресію, але у нас немає програмного забезпечення для цього. Натомість ми маємо лише програмне забезпечення для оцінки простої лінійної регресії. Тому ми використаємо наступний ітераційний підхід: ми фіксуємо всі оцінки коефіцієнтів, крім одного на поточному значенні та оновлюємо лише одну оцінку коефіцієнта за допомогою простої лінійної регресії. Процес продовжується поки не досягнеться збіжність, тобто поки оцінки коефіцієнтів не перестануть змінюватися. Застосуємо цей підхід на штучному прикладі.

6.1. Згенеруйте залежну змінну *Y* і два предиктори *X*1 і *X*2, з *n* = 100.

6.2. Ініціалізуйте оцінку *β*1 довільним значенням на свій вибір.

6.3. Не змінюючи *β*1 оцініть модель

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

Це можна зробити наступним чином:

> a=y-beta1 \*x1

> beta2=lm(a*∼*x2)$coef [2]

6.4. Зафіксувавши оцінку *β*2, оцініть модель

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

Це можна зробити наступним чином:

> a=y-beta2 \*x2

> beta1=lm(a*∼*x1)$coef [2]

6.5. Використайте for для організації циклу з повторень кроків 6.3 та 6.4 1,000 разів. Наведіть оцінки параметрів *β*0, *β*1 і *β*2 на кожній ітерації. Побудуйте графіки на яких відображено ці значення для *β*0, *β*1 і *β*2 різними кольорами.

6.6. Порівняйте результати 6.5 з результатами оцінки множинної регресії *Y* на *X*1 і *X*2. Використайте функцію abline() для накладання цих значень на графік отриманий в 6.5.

6.7. Для цього набору даних, скільки ітерацій підгонки потрібно було для отримання «доброго» наближення до оцінок коефіцієнтів множинної регресії?

7. Продовжимо розгляд попереднього прикладу. Покажіть, що у випадку *p* = 100 можна отримати оцінки коефіцієнтів множинної регресії повторно застосовуючи метод підгонки. Скільки ітерацій підгонки потрібно було для отримання «доброго» наближення до оцінок коефіцієнтів множинної регресії? Побудуйте графік для підтвердження своїх висновків.