1. Проста лінійна регресія на основі даних Auto.

1.1. Використайте функцію lm() для побудови простої лінійної регресії з залежною змінною mpg і незалежною – horsepower. Виведіть та прокоментуйте результати з допомогою summary(). Наприклад,

Чи існує взаємозв'язок між предиктором та залежною змінною? Наскільки сильний зв’язок між предиктором та залежною змінною? Чи взаємозв'язок між предиктором та залежною змінною позитивний чи негативний? iv. Яке прогнозне значення залежної змінної при значенні предиктора 98? Які пов'язані 95% інтервали довіри та прогнозування?

1.2. Зобразіть графічно предиктор та залежну змінну. Використайте функцію abline() для відображення оціненої прямої.

1.3. Використайте функцію plot() для зображення діагностичних графіків. Прокоментуйте які ви бачите проблеми.

2. Множинна лінійна регресія на основі даних Auto.

2.1. Побудуйте діаграми розкиду усіх змінних.

2.2. Обчисліть матрицю кореляцій між змінними використовуючи функцію cor().

2.3. Використовуючи функцію lm() побудуйте множинну регресію для залежної змінної mpg і всіх решту змінних окрім names як предикторів. На основі summary() прокоментуйте результати. Наприклад:

Чи існує взаємозв'язок між предикторами та залежною змінною? Які предиктори мають статистично значущі показники зв’язку з залежною змінною? iii. Що означає коефіцієнт для year?

2.4. Використайте функцію plot () для створення діагностичних графіків. Прокоментуйте будь-які проблеми, які ви бачите. Чи вказують графіки на наявність великих викидів? Чи вказують графіки на наявність спостережень з високим левереджем?

2.5. Використовуйте символи \* та:, щоб включити в модель лінійної регресії ефекти взаємодії. Чи є розглянуті змінні взаємодії статистично значимі?

2.6. Спробуйте кілька різних перетворень змінних, таких як log(X), *X*2, .



3. Розглянемо дані Carseats.

3.1. Побудуйте модель множинної регресії для прогнозування Sales використовуючи Price, Urban, та US..

3.2. Надайте інтерпретацію кожного коефіцієнта в моделі.

3.3. Запишіть модель у формі рівняння.

3.4. Для якого з предикторів можна відхилити нульову гіпотезу H0: βj = 0?

3.5. На основі відповіді на попереднє запитання побудуйте модель з меншою кількістю незалежних змінних, яка використовує лише ті предиктори, для яких зв’язок з залежною змінною є значимим.

3.6. Яка з попередніх моделей є краща?

3.7. Використовуючи модель з 3.5, побудуйте 95% інтервали довіри для коефіцієнтів.

3.8. Чи можна сказати, що присутніми є викиди та спостереження з високим рівнем левереджу для моделі з 3.5?

4. У цій задачі ми дослідимо *t*-статистику для нульової гіпотези H0: *β* = 0 у простій лінійній регресії без коефіцієнта *β*0. Для початку ми згенеруємо предиктор *x* та залежну змінну *y*.

> set.seed (1)

> x=rnorm (100)

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

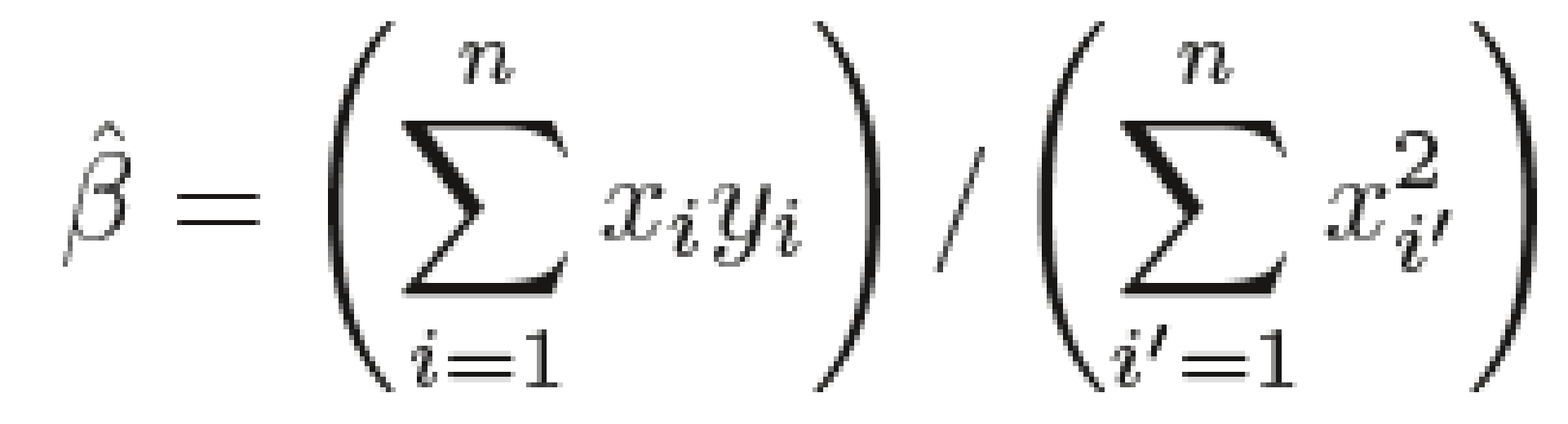
4.1. Побудуйте просту лінійну регресію *y* на *x* без *β*0. Наведіть значення оцінки коефіцієнта *β*, його стандартну похибку, *t*-статистику та *р*-значення пов'язані з нульовою гіпотезою H0: *β* = 0. Прокоментуйте ці результати. (Побудувати регресію без *β*0 можна за допомогою команди lm (y∼x + 0).)

4.2 Побудуйте просту лінійну регресію *x* на *y* без *β*0. Наведіть значення оцінки коефіцієнта *β*, його стандартну похибку, *t*-статистику та *р*-значення пов'язані з нульовою гіпотезою H0: *β* = 0. Прокоментуйте ці результати.

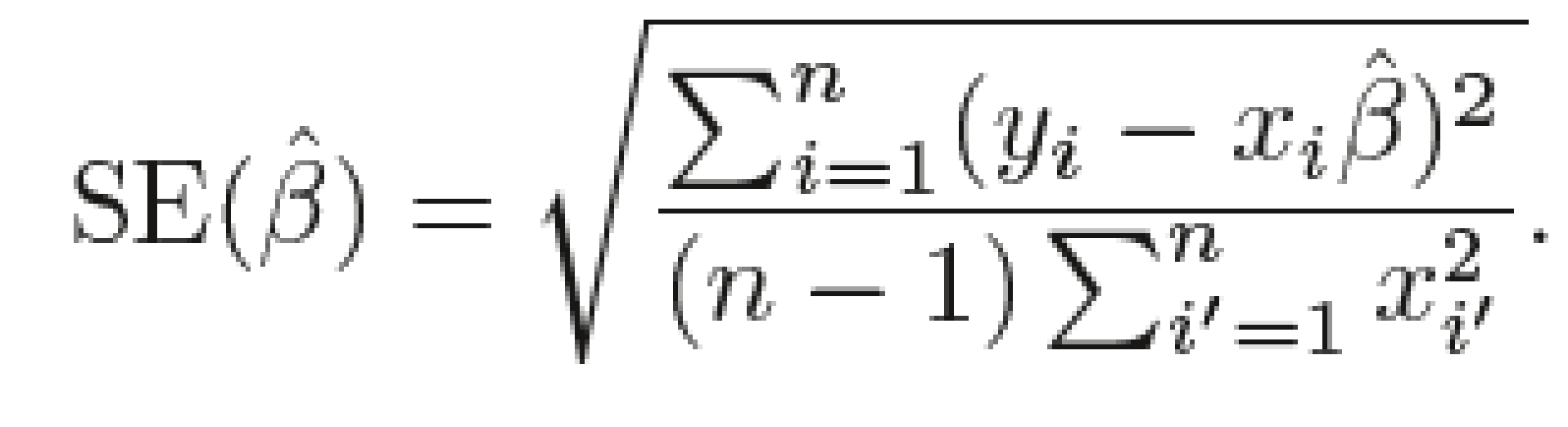
4.3. Який зв’язок між результатами, отриманими в 4.1 та 4.2?

4.4. Для регресії Y на X без *β*0, статистика для H0: *β* = 0 приймає вигляд , де

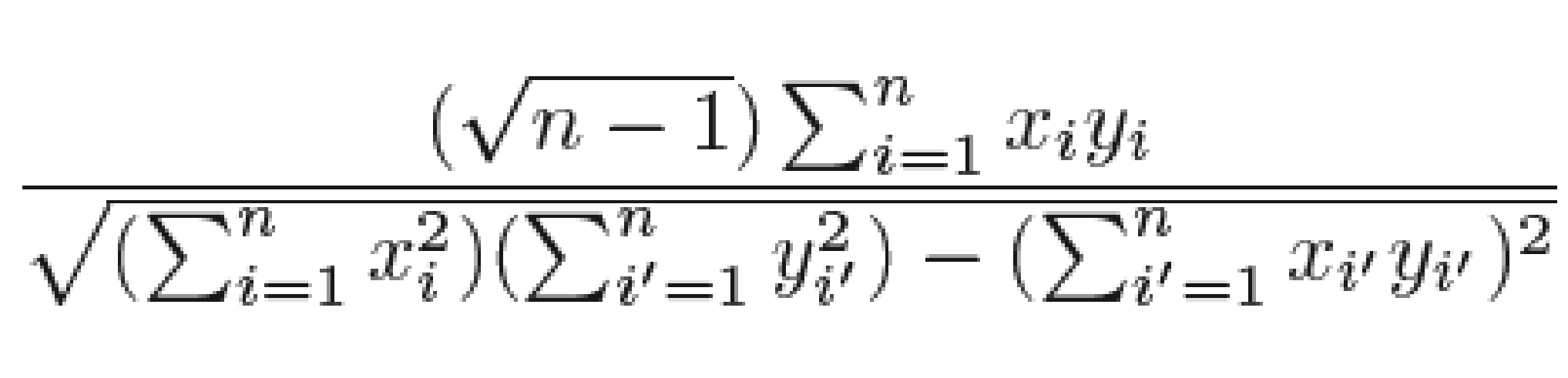




та



Доведіть та перевірте чисельно, що t-статистика може бути записана у вигляді

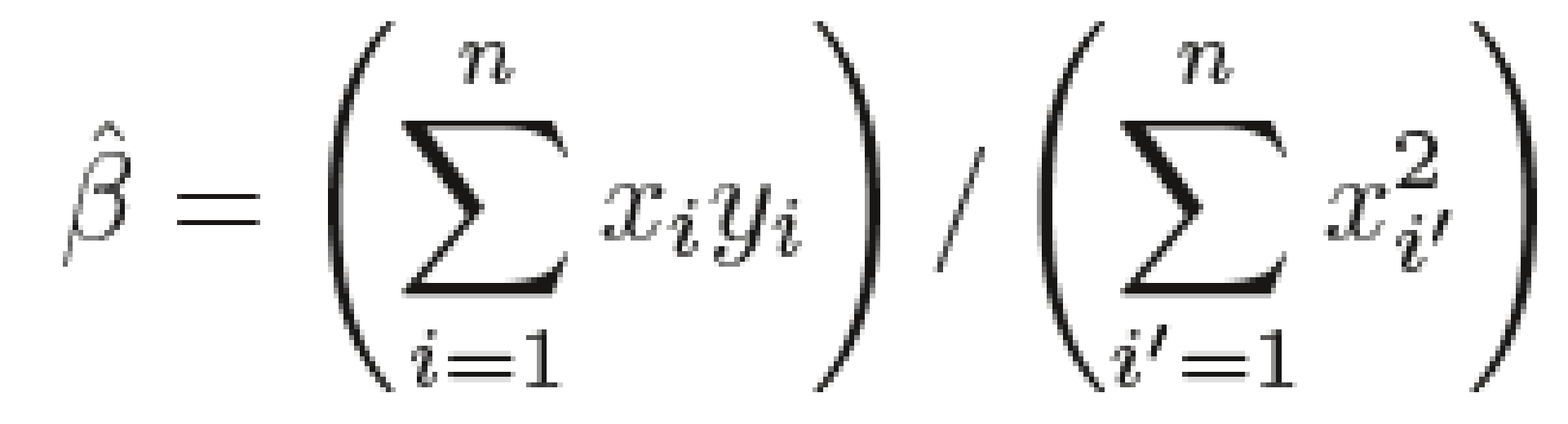


4.5. Використовуючи результати з 4.4, обгрунтуйте, що *t*-статистика для регресії *y* на *x* є те саме, що *t*-статистика для регресії *x* на *y*.

4.6. Покажіть, що коли будується регресія з коефіцієнтом *β*0, то *t*-статистика для H0: *β*1 = 0 однакова для регресії *y* на *x*, та для регресії *x* на *y*.

5. Знову розглянемо просту лінійну регресію без коефіцієнта *β*0.

5.1. Нагадаємо, що оцінка коефіцієнта β для лінійної регресії Y на X без коефіцієнта *β*0 має вигляд:



В якому випадку оцінка коефіцієнта регресії X на Y дорівнює оцінці коефіцієнта регресії Y на X?

5.2. Побудуйте приклад у R з n = 100 спостережень, в якому оцінка коефіцієнта для регресії X на Y не дорівнює оцінці коефіцієнта регресії Y на X.

5.3. Побудуйте приклад у R з n = 100 спостережень, в якому оцінка коефіцієнта для регресії X на Y дорівнює оцінці коефіцієнта регресії Y на X.

6. У цій вправі потрібно згенерувати набір даних та оцінити кілька простих лінійних моделей. Використайте set.seed (1) перед початком частини 6.1 для забезпечення однакових результатів.

6.1 За допомогою функції rnorm () створіть вектор x, що містить 100 спостережень, отриманих з розподілу N(0, 1).

6.2. За допомогою функції rnorm() створіть вектор eps, що містить 100 спостереження, отриманих з розподілу N(0, 0,25).

6.3. За допомогою x та eps згенеруйте вектор y відповідно до моделі

Y = – 1 + 0,5X + *ε*

Яка довжина вектора y? Які значення *β*0 і *β*1 у цій лінійній моделі?

6.4. Побудуйте діаграму розсіювання, що відображає взаємозв'язок між x та y. Прокоментуйте результати.

6.5. Побудуйте лінійну модель для прогнозування y на основі x. Прокоментуйте на отриманій моделі як оцінки параметрів *β*0 та *β*1 співвідносяться з точними значеннями цих параметрів?

6.6. Відобразіть оцінену лінію на діаграмі розсіяння, отриманій в 6.4. Накресліть цю лінію іншим кольором. За допомогою команди legend () створіть відповідну легенду.

6.7. Побудуйте модель поліноміальної регресії, яка передбачає y на основі x і x2. Чи можна сказати, що квадратичний доданок покращує модель? Поясніть свою відповідь.

6.8. Повторіть 6.1. – 6.6. після модифікації процесу генерації даних *у* таким чином, щоб було менше шуму в даних. Модель для *y* повинна залишатися незмінною. Це можна зробити, зменшивши дисперсію нормального розподілу, що використовується для генерування залишків. Прокоментуйте результати.

6.9. Повторіть 6.1. – 6.6. після модифікації процесу генерації даних *у* таким чином, щоб було більше шуму в даних. Модель для *y* повинна залишатися незмінною. Це можна зробити, збільшивши дисперсію нормального розподілу, що використовується для генерування залишків. Прокоментуйте результати.

6.10. Які довірчі інтервали для *β*0 та *β*1 на основі оригінальних даних, даних з більшим шумом та даних з меншим шумом? Обгрунтуйте.

7. Зосередимося на проблемі колінеарності.

7.1. Виконайте такі команди в R:

> set .seed (1)

> x1=runif (100)

> x2 =0.5\* x1+rnorm (100) /10

> y=2+2\* x1 +0.3\* x2+rnorm (100)

Останній рядок відповідає створенню лінійної моделі, в якій *y* є функція від x1 і x2. Випишіть форму лінійної моделі. Які коефіцієнти регресії?

7.2. Яка кореляція між x1 та x2? Побудуйте діаграму розсіювання для відображення зв'язку між змінними.

7.3. Використовуючи ці дані, оцініть регресію методом найменших квадратів, щоб передбачити *y*, використовуючи x1 та x2. Опишіть отримані результати. Як оцінки параметрів відносяться до точних значень? Чи можна відхилити нульову гіпотезу H0: *β*1 = 0? Як щодо гіпотези H0: *β*2 = 0?

7.4. Побудуйте регресію y на x1. Прокоментуйте свої результати. Чи можна відхилити нульову гіпотезу H0: *β*1 = 0?

7.5. Побудуйте регресію y на x2. Прокоментуйте свої результати. Чи можна відхилити нульову гіпотезу H0: *β*2 = 0?

7.6. Чи суперечать результати, отримані в 7.3 – 7.5 один одному? Поясніть вашу відповідь.

7.7. Припустимо, тепер ми отримуємо одне додаткове спостереження, яке було на жаль, неправильно виміряно.

> x1 = c (x1, 0,1)

> x2 = c (x2, 0,8)

> y = c (y, 6)

Переоцініть попередні лінійні моделі, використовуючи ці нові дані. Чи впливає це нове спостереження на кожну з моделей? Чи у кожній моделі це спостереження є викидом? З високим рівнем левереджу? І те, і інше? Поясніть свої відповіді.

8. Розглянемо набір даних Boston. Ми спробуємо спрогнозувати рівень злочинності на душу населення використовуючи інші змінні в цьому наборі даних. Іншими словами, рівень злочинності на душу населення – залежна змінна, а інші змінні - предиктори.

8.1. Для кожного предиктора побудуйте просту модель лінійної регресії для прогнозування рівня злочинності на душу населення. Опишіть свої результати. В якій з моделей існує статистично значущий зв’язок між предиктором та залежною змінною? Побудуйте кілька графіків для більшої наочності.

8.2. Побудуйте модель множинної регресії для прогнозування залежної змінної за допомогою всіх предикторів. Опишіть свої результати. Для яких предикторів ми можемо відхилити нульову гіпотезу H0: *βj* = 0?

8.3. Як ваші результати з 8.1 співвідносяться з результатами з 8.2?

8.4. Чи є ознаки нелінійності зв’язку між будь-якими з предикторів та залежною змінною? Щоб відповісти на це питання, для кожного предиктора X, побудуйте модель вигляду

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