МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ЛЬВІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ІВАНА ФРАНКА

Факультет прикладної математики та інформатики

ЗВІТ

до індивідуального завдання №2

з дисципліни «Моделі статистичного навчання»

Виконали

студенти групи ПМіМ-12:

Бордун Михайло

Зелінський Олександр

Перевірив:

Проф. Заболоцький Т. М.

Львів – 2021

**Хід виконання**

**1. Проста лінійна регресія на основі даних Auto**

**1.1**

Зображення, що містить текст

Автоматично згенерований опис

Зображення, що містить текст

Автоматично згенерований опис

Так, існує залежність між horsepower та mpg, яка визначена шляхом перевірки нульової гіпотези всіх коефіцієнтів регресії, рівних нулю. Оскільки F-статистика набагато більша за 1, а p-значення F-статистики близьке до нуля, ми можемо відкинути нульову гіпотезу і стверджувати, що існує статистично значуща залежність між horsepower та mpg.

Співвідношення між mpg і horsepower є негативним. Чим більше horsepower в автомобіля, тим меншою є mpg автомобіля.

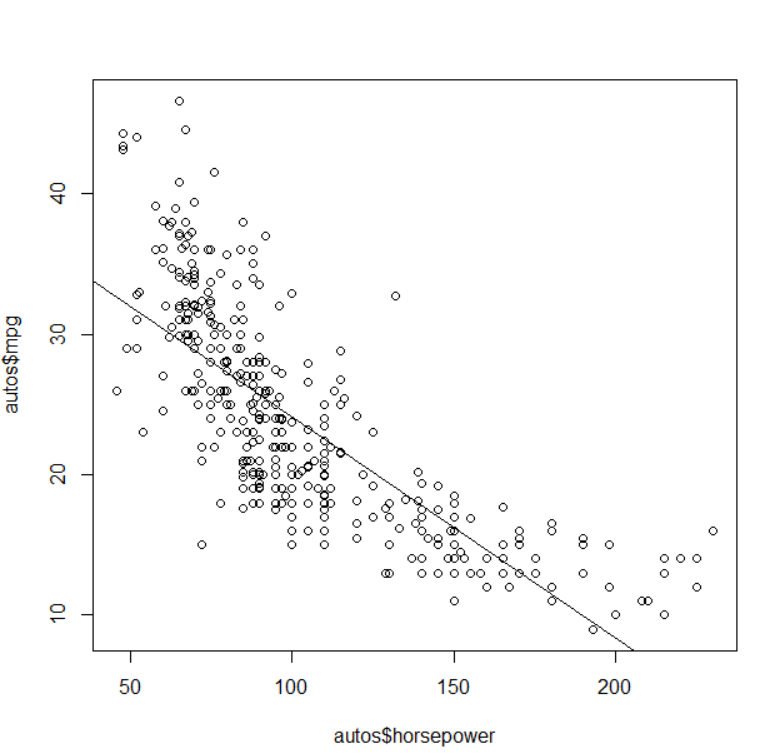
Зображення, що містить текст

Автоматично згенерований опис

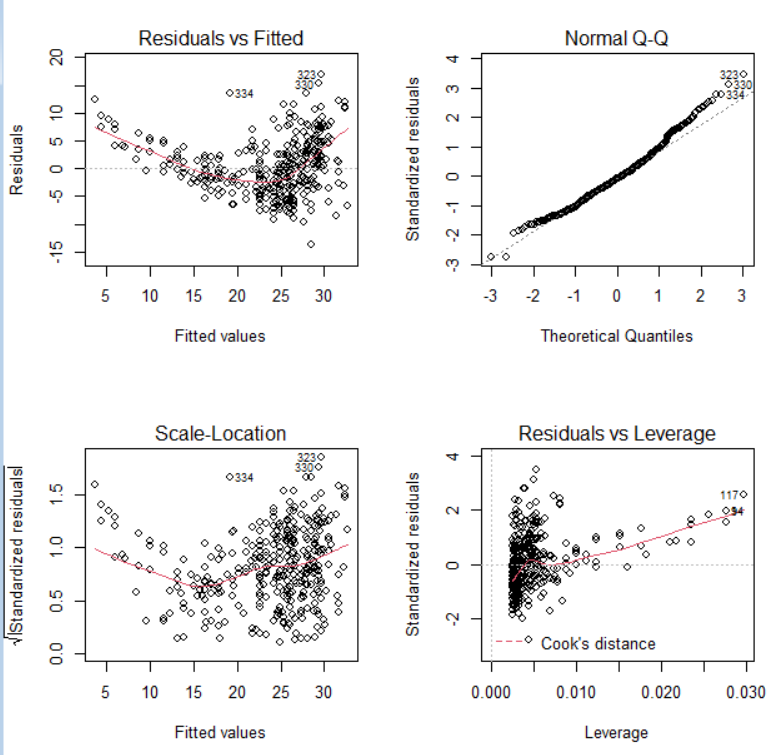
Зображення, що містить текст

Автоматично згенерований опис

**1.2**



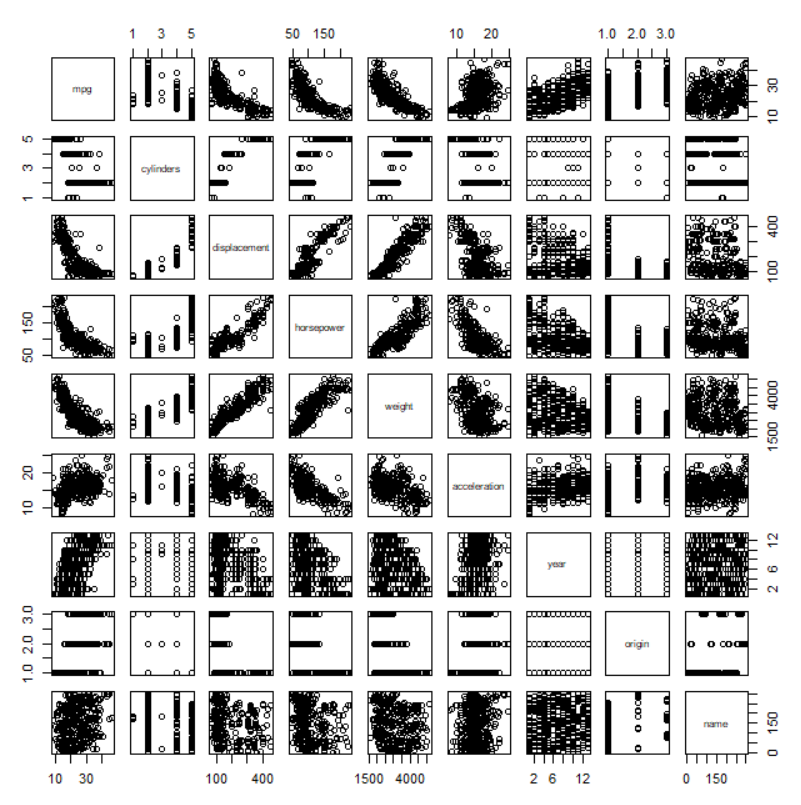
**1.3**



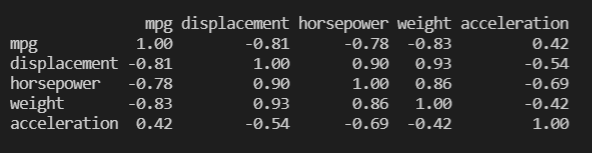
Зважаючи на ці графіки можна сказати, що залежність не зовсім лінійна.

**2. Множинна лінійна регресія на основі даних Auto.**

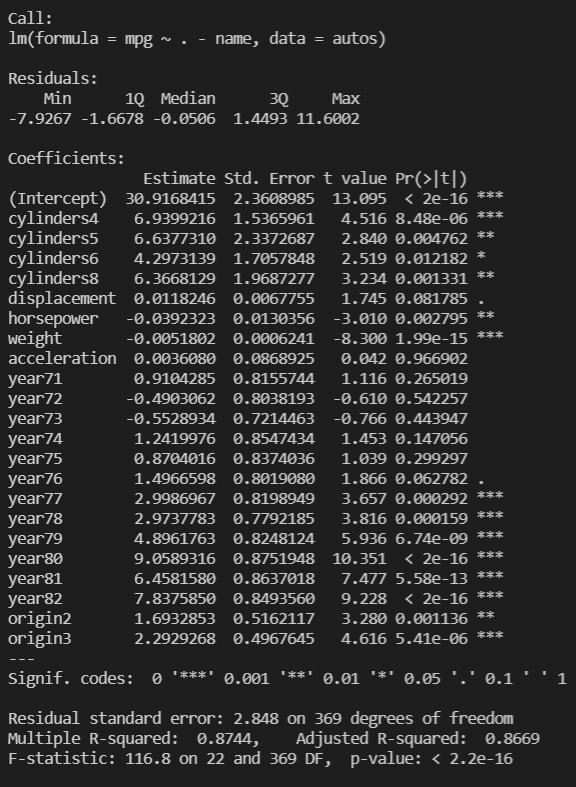
**2.1**

****

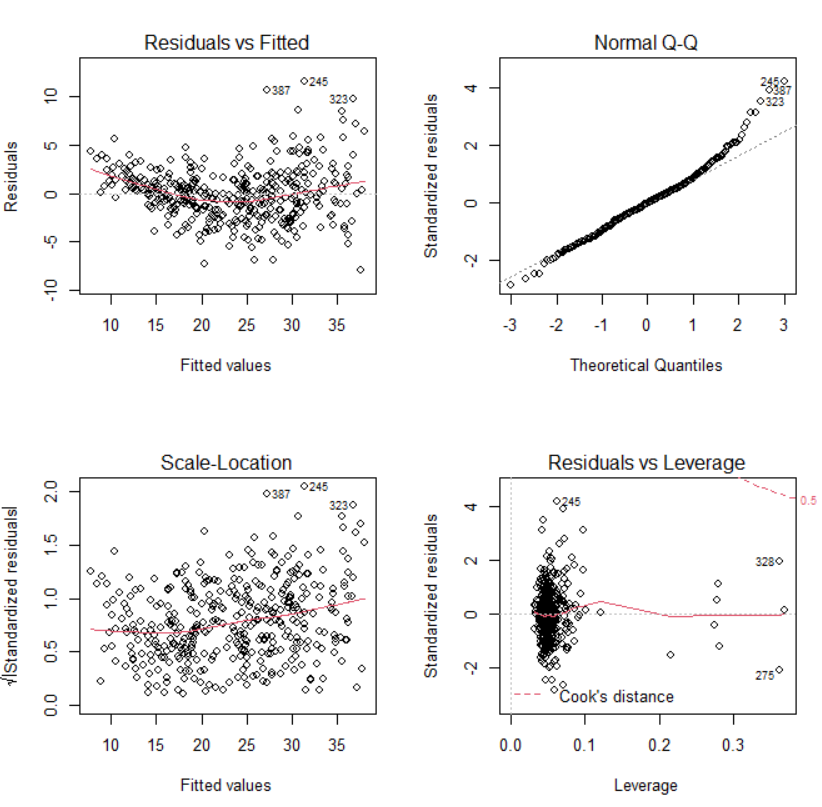
**2.2**

****

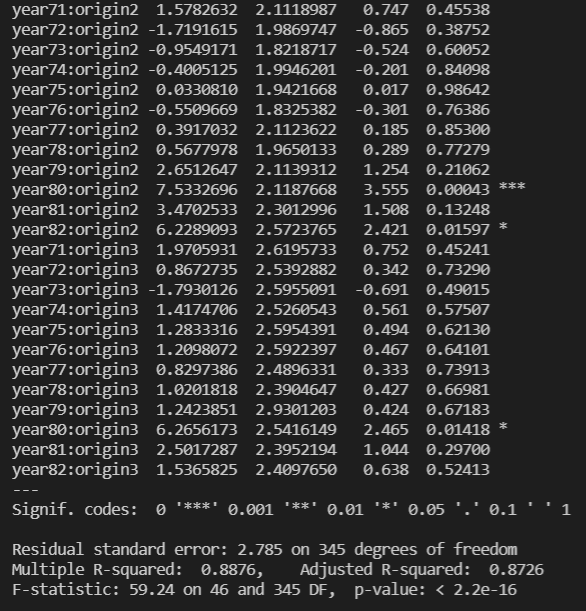
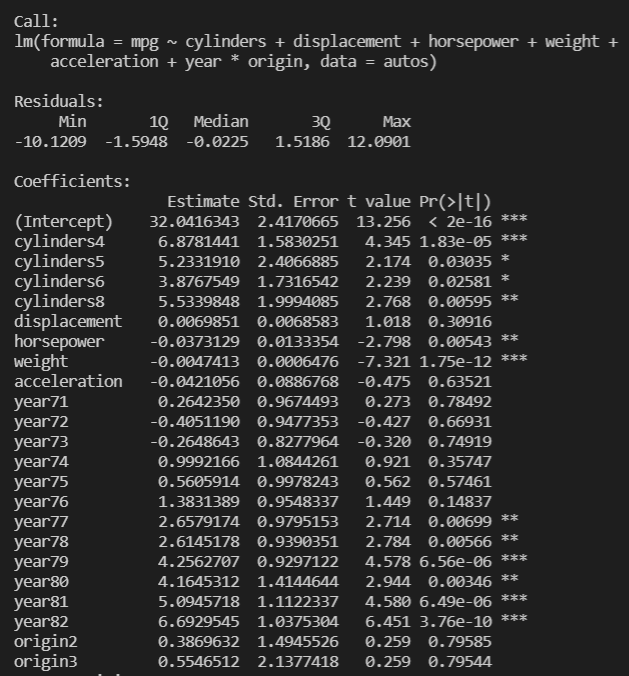
**2.3**

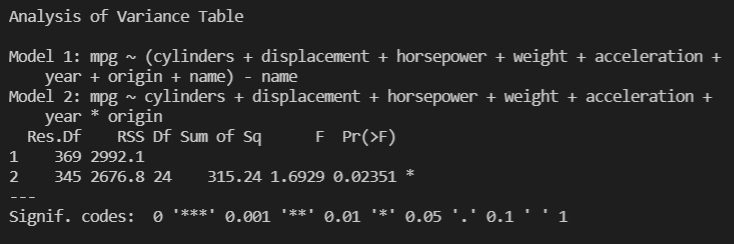
****

**2.4**

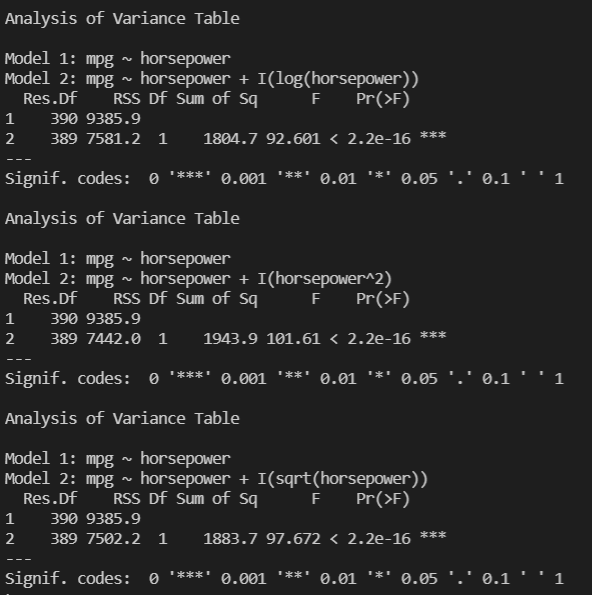
****

**2.5**

****

****

**2.6**

****

**3. Розглянемо дані Carseats.**

Можемо побачити дані та їх опис.

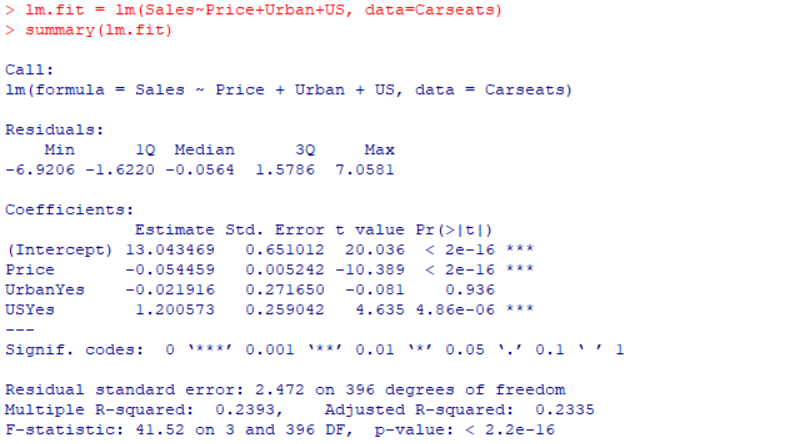
**Зображення, що містить стіл

Автоматично згенерований опис**

Зображення, що містить текст

Автоматично згенерований опис

3.1



**3.2**

* **Price.** Лінійна регресія передбачає зв’язок між Price та Sales з огляду на низьку р-величину t-статистики. Коефіцієнт свідчить про негативне співвідношення між Price та Sales: із зростанням Price, Sales зменшується.
* **UrbanYes.** Лінійна регресія свідчить про відсутність залежності між місцем розташування магазину та кількістю продажів на основі високої p-вартості t-статистики.
* **USYes.** Лінійна регресія свідчить про існування залежності між тим, чи знаходиться магазин у США чи ні, та обсягом продажів. Коефіцієнт свідчить про позитивне співвідношення між USYes та Sales: якщо магазин знаходиться в США, продажі збільшаться приблизно на 1201 одиницю.

**3.3**

**3.4**

Нульову гіпотезу можна відхилити для гіпотези Price та USYes, на основі колонки .

**3.5** Нова модель

Зображення, що містить текст

Автоматично згенерований опис

**3.6**

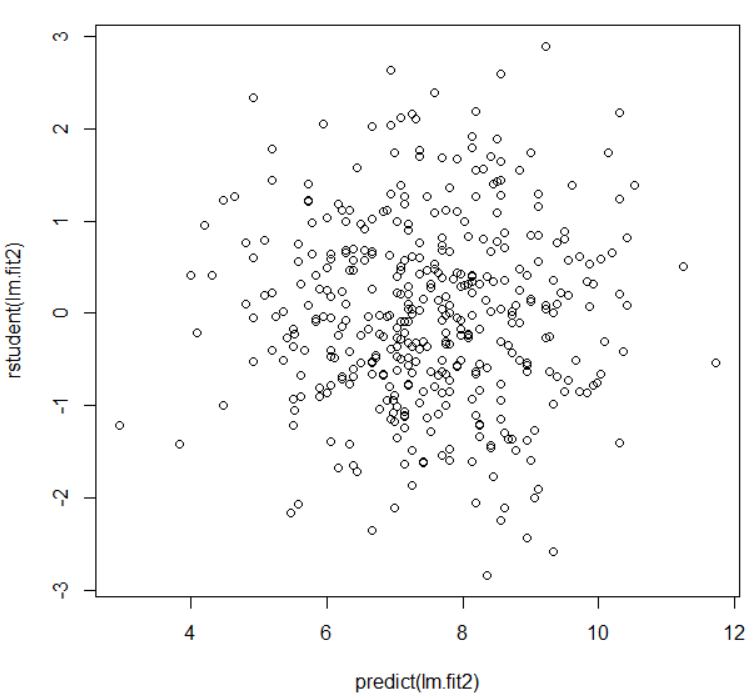
Зважаючи на значення RSE та можна стверджувати, що обидві моделі добре підходять для даних. Проте друга модель трошки краща.

**3.7**

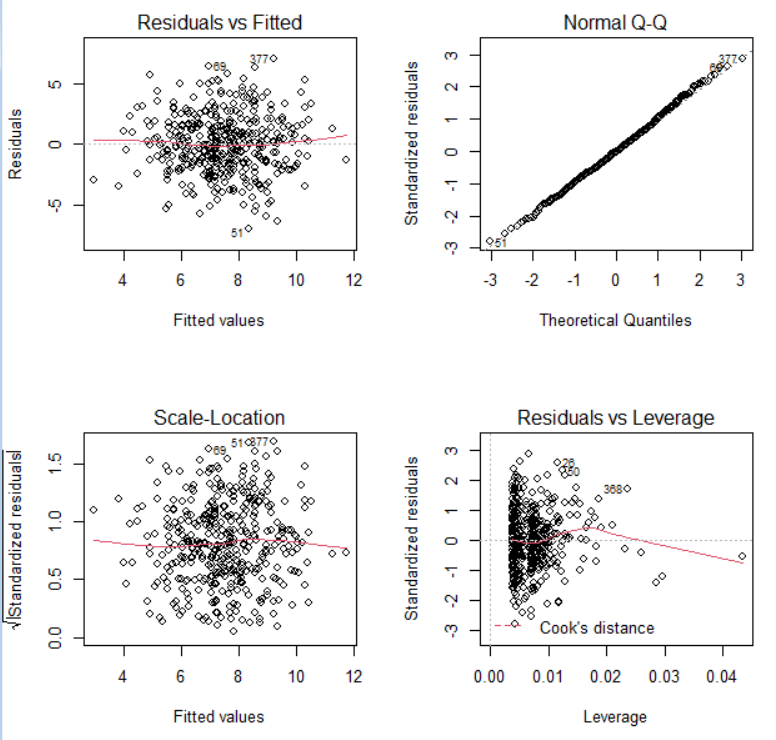
Зображення, що містить текст

Автоматично згенерований опис

**3.8**

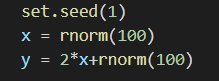


Усі Стюдентифіковані залишки, обмежені від -3 до 3, тому з лінійної регресії не випливають потенційні викиди.

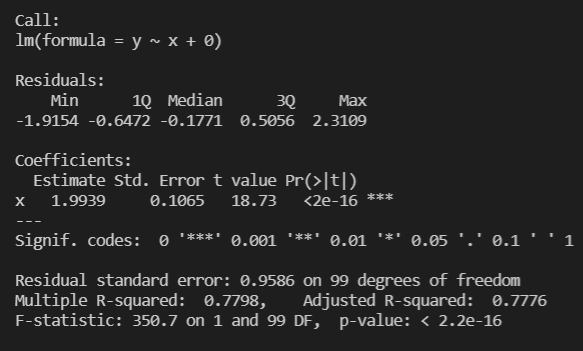


Існує декілька спостережень, які значно перевищують (p+1)/n (0,0076) на графіку leverage-statistic, що свідчить про те, що відповідні точки мають високий leverage.

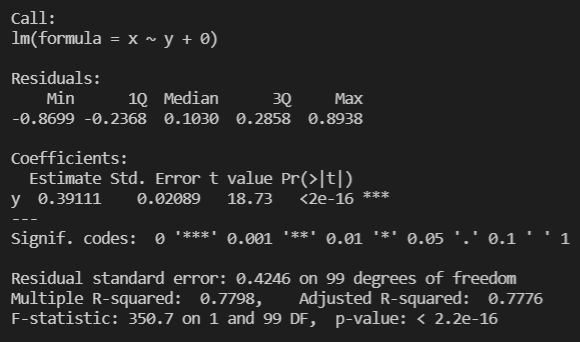
**4. Дослідження t-статистики для нульової гіпотези у простій лінійній регресії без коефіцієнта *β*0.**



**4.1**

****

**4.2**

****

**4.3**

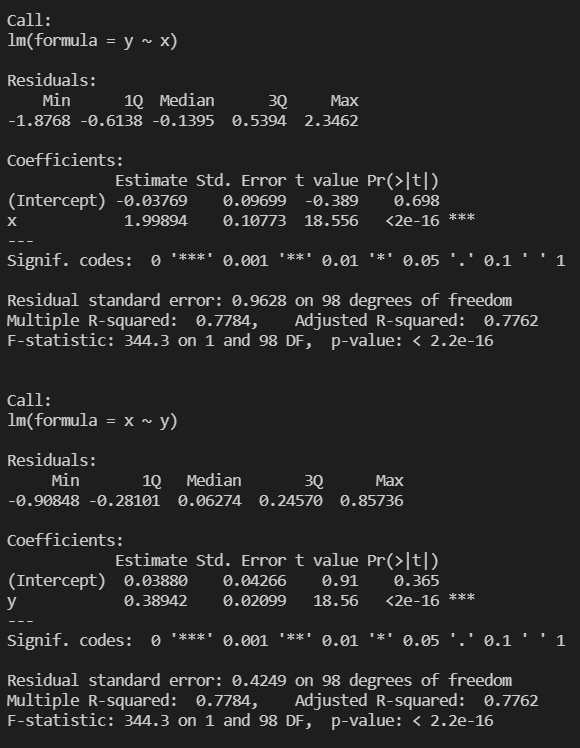
****

**4.4**

****

**4.5**

**4.6**

****

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

**5.1**

Коефіцієнт регресії X на Y буде рівним оцінці коефіцієнта регресії Y на X коли:

**5.2** Створимо вектори x та y з різними сумами квадратів їх елементів

Зображення, що містить текст

Автоматично згенерований опис

Після цього оцінимо коефіцієнти для лінійної регресії Y на X та X на Y

Зображення, що містить текст

Автоматично згенерований опис

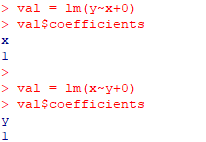
З результатів видно що вони різні

**5.3** Згенеруємо вектори x та y, такі щоб суми квадратів їх елементів були рівними.

Зображення, що містить текст

Автоматично згенерований опис

З результатів можна побачити, що коефіцієнти для лінійної регресії Y на X та X на Y однакові.



**6. Генерування набору даних та оцінка кількох простих лінійних моделей.**

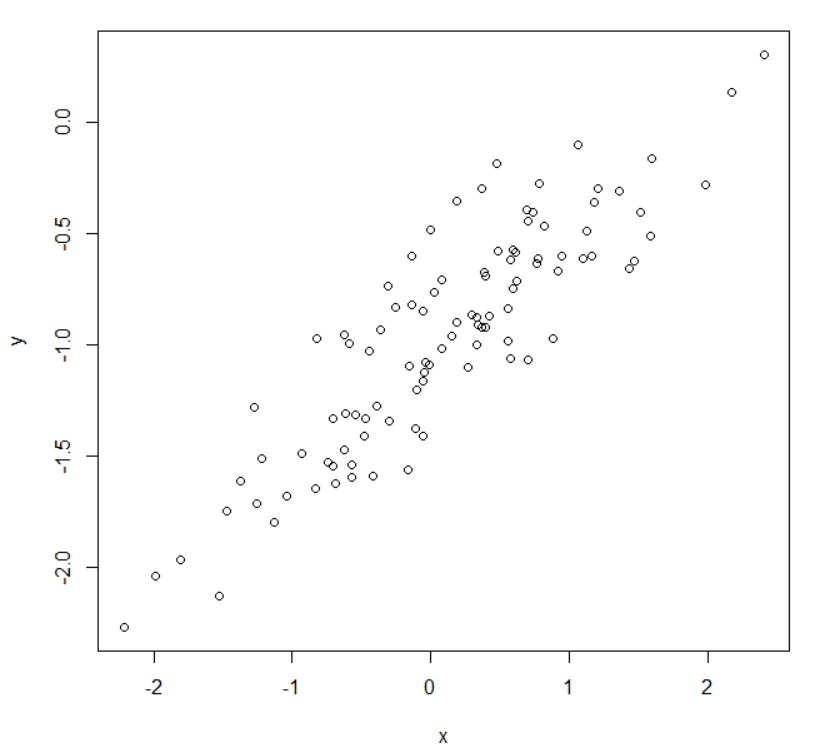
**6.1**

**6.2**

**6.3**

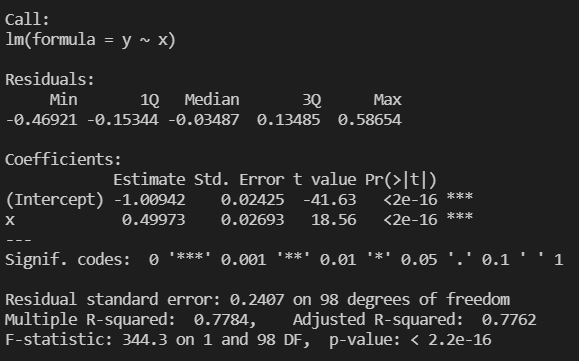
****

**6.4**

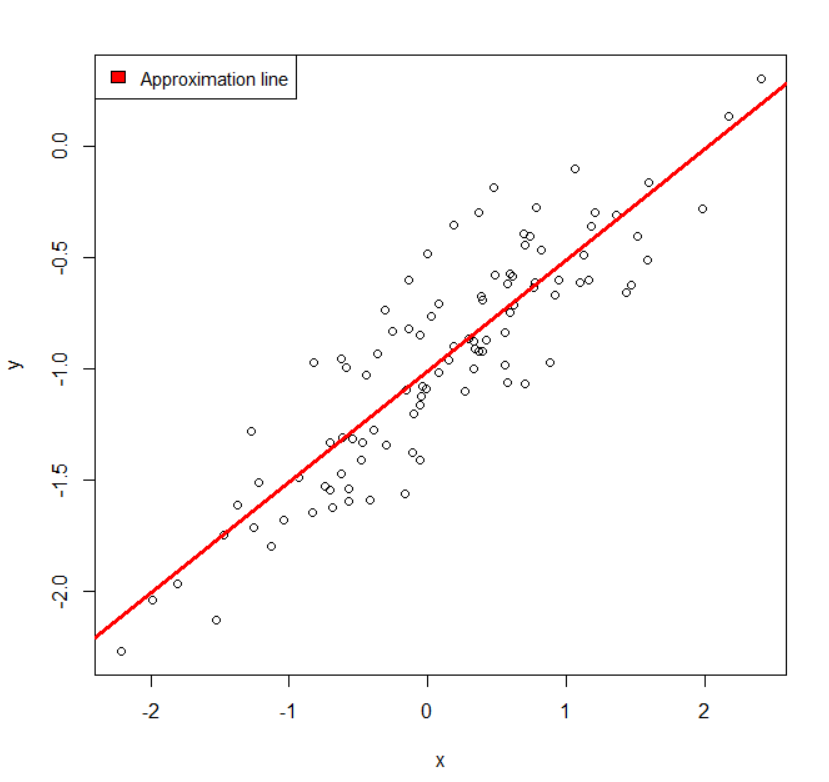
****

****

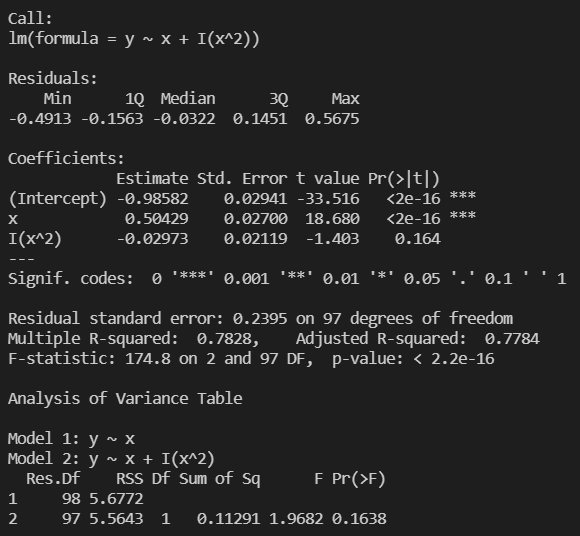
**6.5**

****

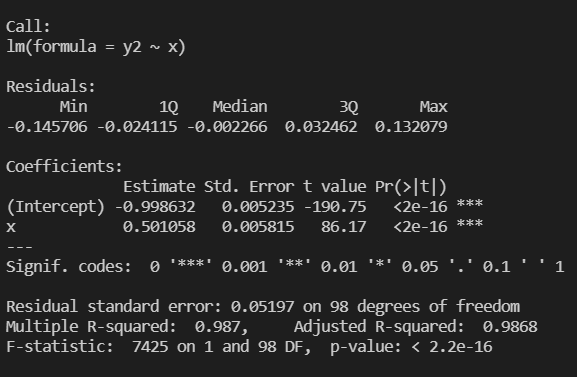
**6.6**

****

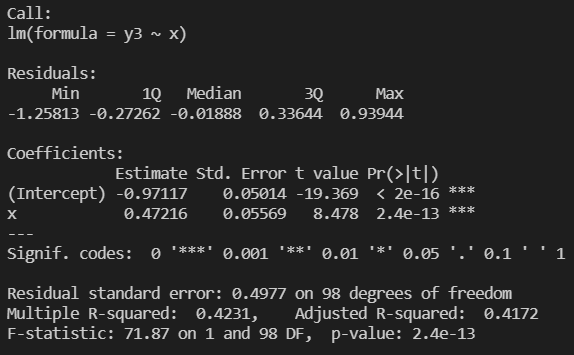
**6.7**

****

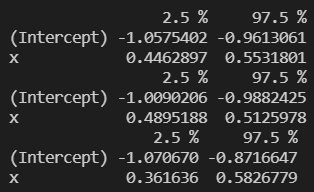
**6.8**

****

**6.9**

****

**6.10**

****

**7. Зосередимося на проблемі колінеарності.**

**7.1**

Форма лінійної моделі та коефіцієнти регресії

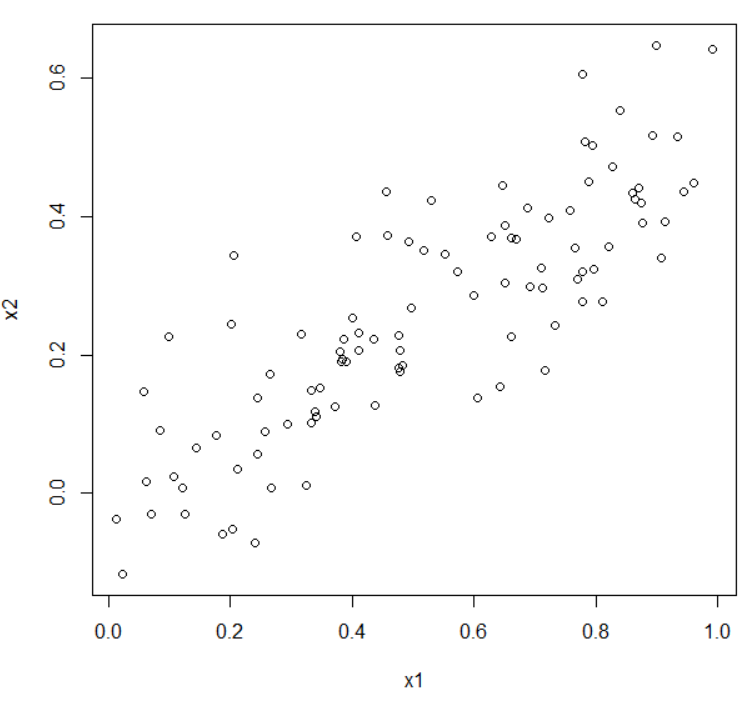
Зображення, що містить текст

Автоматично згенерований опис

**7.2**

****

Кореляція між та та діаграма розсіювання



**7.3**

**7.4**

Зображення, що містить текст

Автоматично згенерований опис

Так, можна відхилити нульову гіпотезу щодо коефіцієнта регресії, бо значення p для його t-статистики близьке до нуля.

**7.5**

Зображення, що містить текст

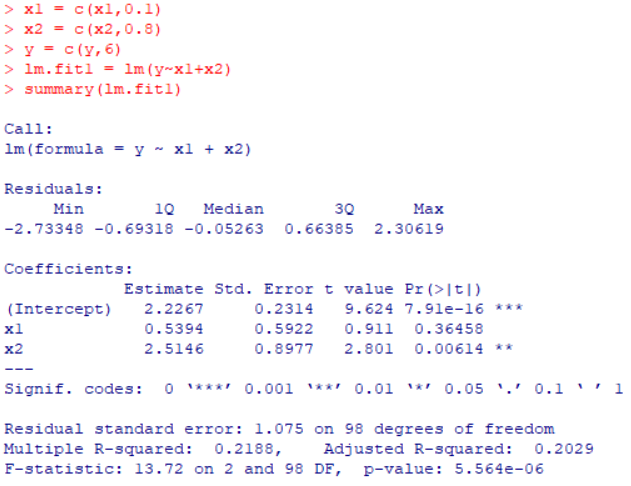
Автоматично згенерований опис

Тут також можна відхилити нульову гіпотезу щодо коефіцієнта регресії, бо значення p для його t-статистики близьке до нуля.

**7.6**

Ні, оскільки та мають колінеарність, важко відрізнити їх вплив, коли вони регресуються разом. Коли вони регресуються окремо, лінійна залежність між і кожним предиктором визначаються більш чітко.

**7.7**



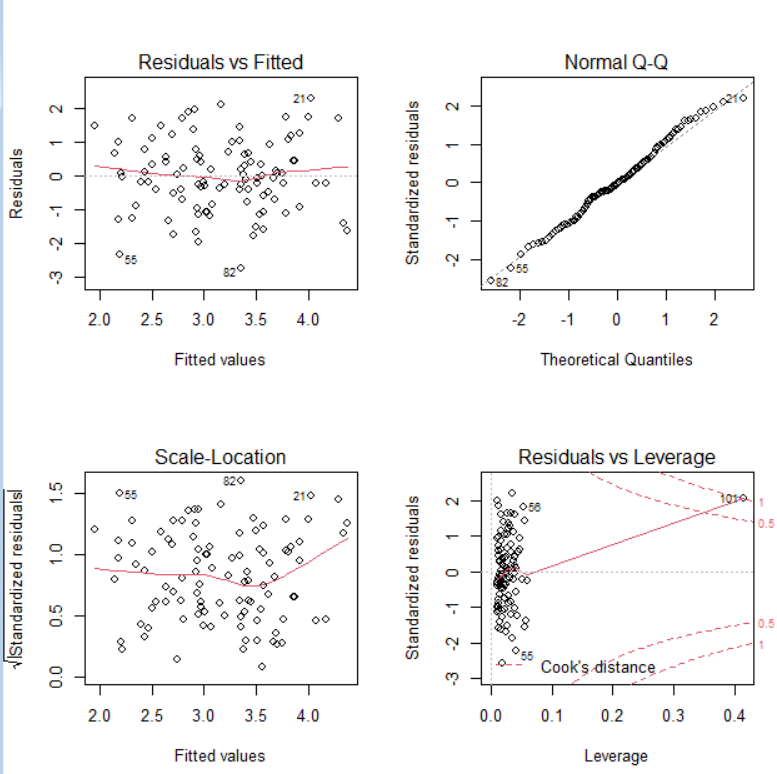
Зображення, що містить текст

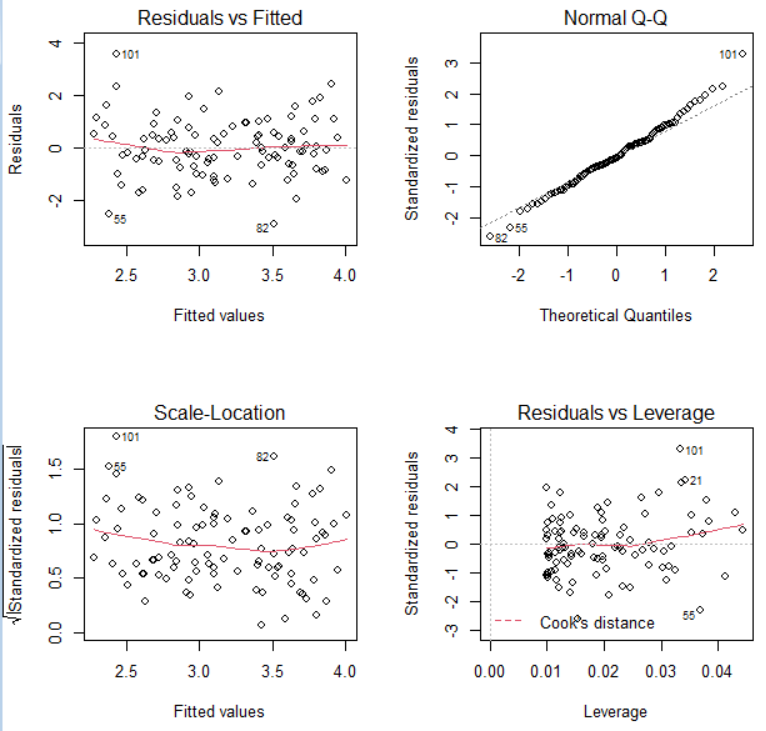
Автоматично згенерований опис

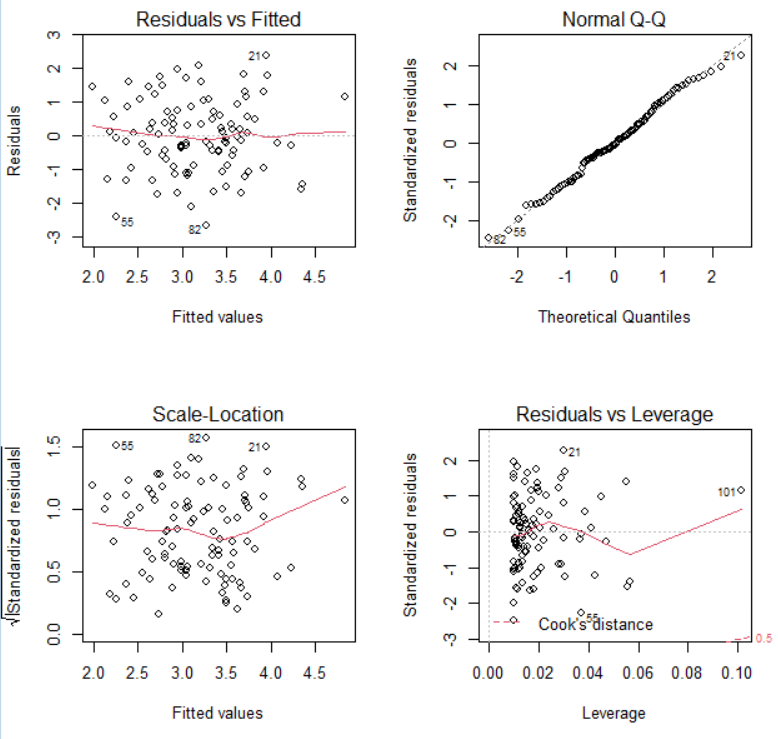
Зображення, що містить текст

Автоматично згенерований опис

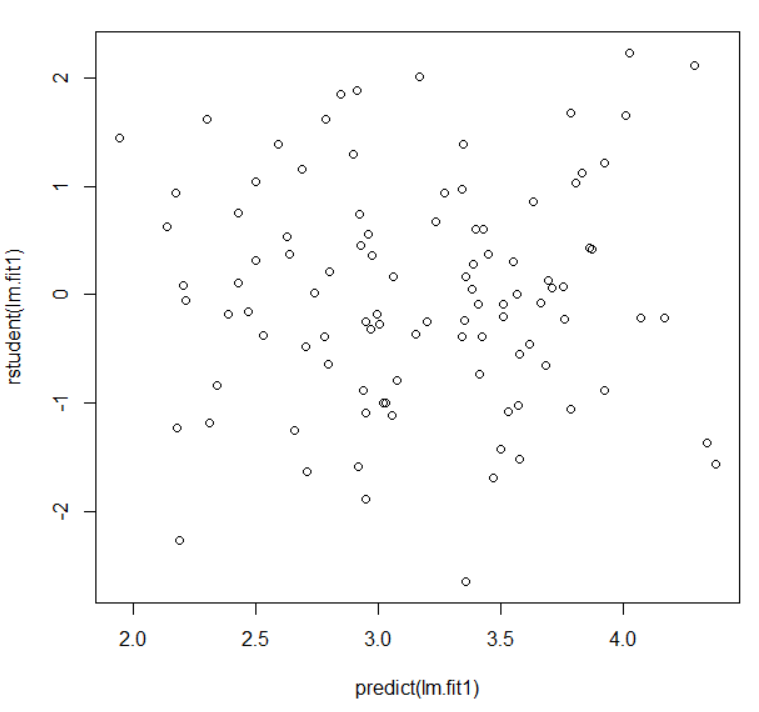
У першій моделі вона зсуває до статистично незначущої та до статистично значущої від зміни p-значень між двома лінійними регресіями.

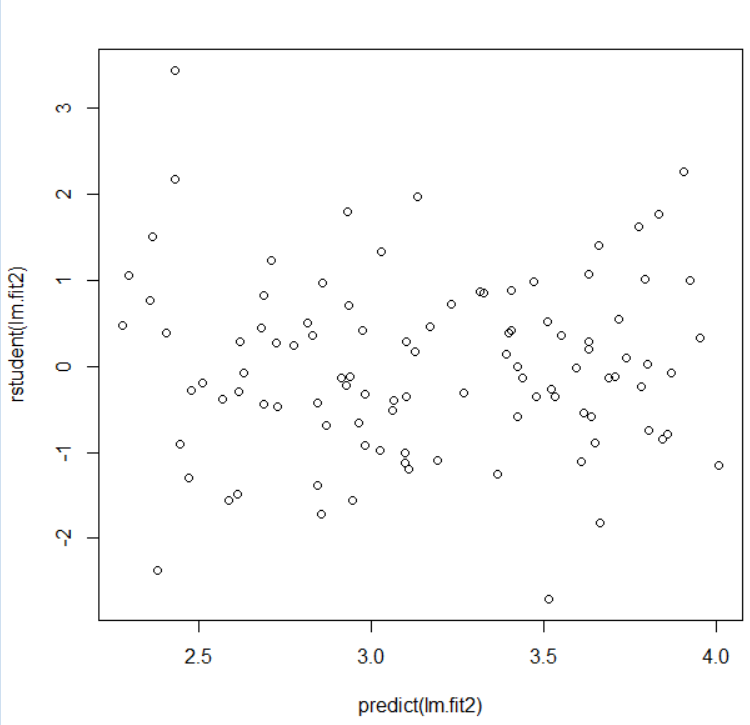


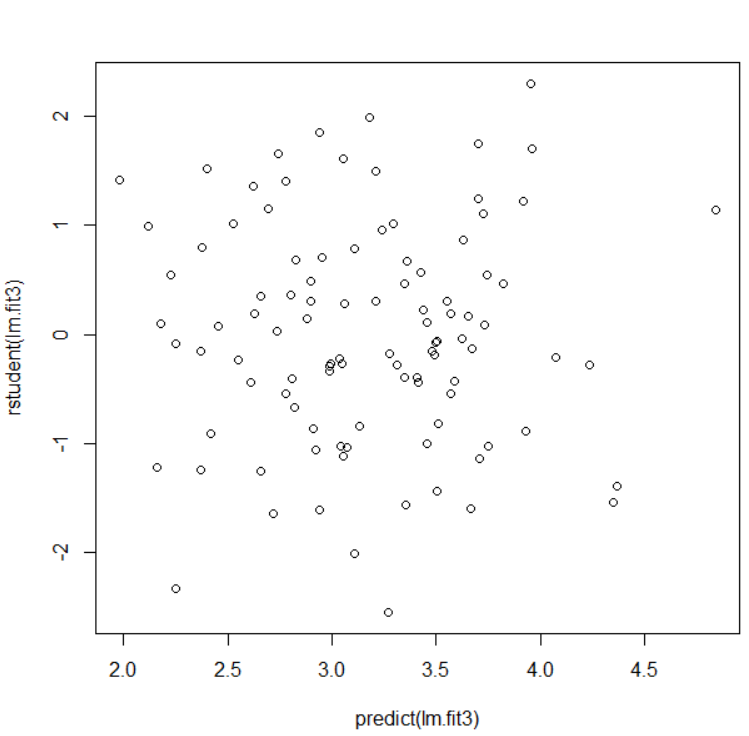




У першій і третій моделях ця точка стає точкою високого leverage.

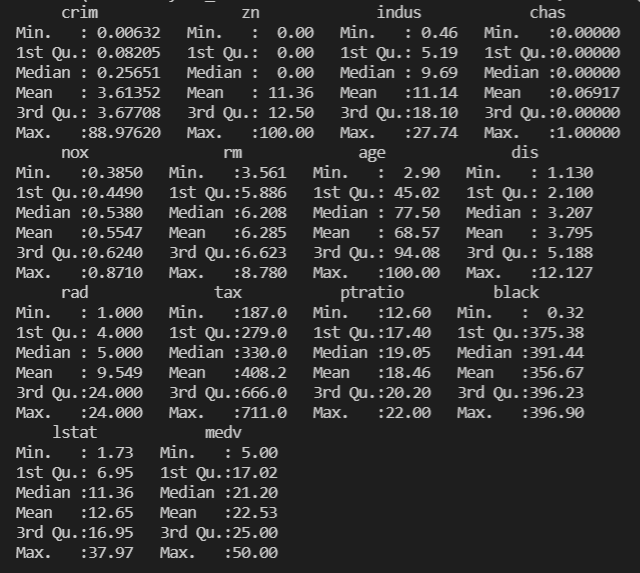






Дивлячись на Стюдентифіковані залишки, ми не спостерігаємо точок занадто далеко від граничного значення, що рівне , за винятком другої лінійної регресії: y ~ x1.

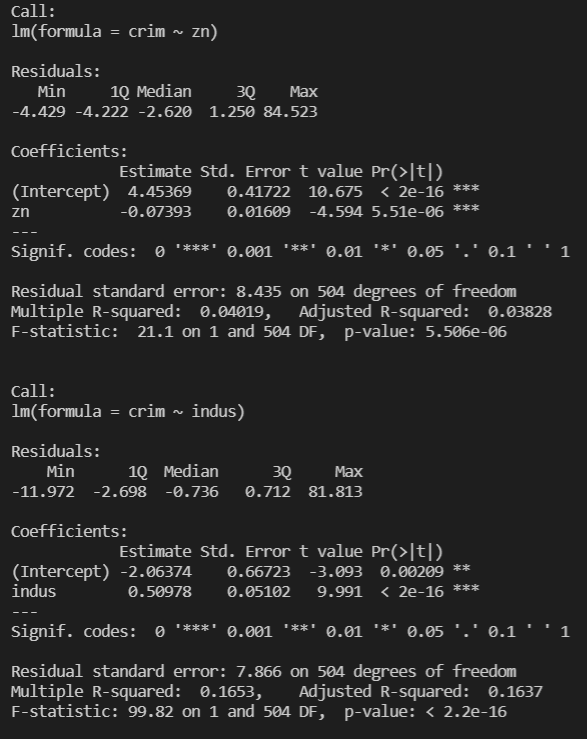
**8. Прогнозування рівня злочинності на душу населення використовуючи інші змінні в наборі даних Boston.**

****

Загальна характеристика даних Boston

**8.1**

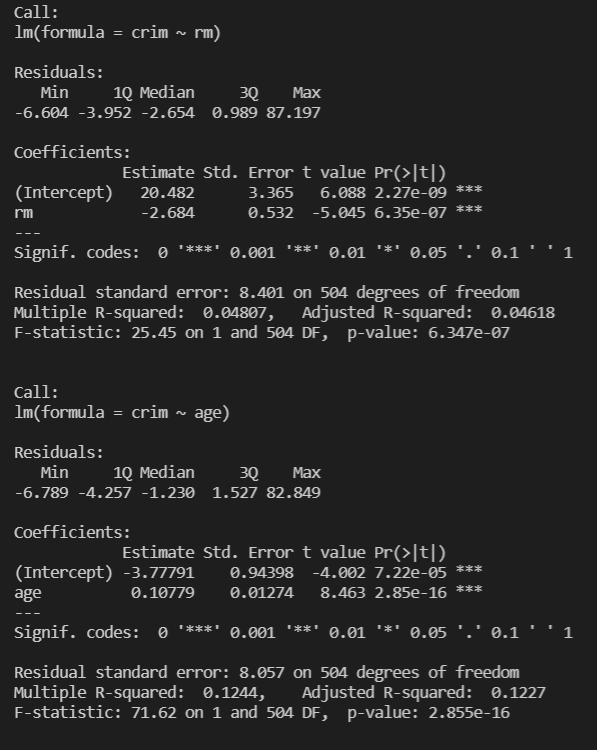
Побудовано для кожного предиктора просту модель лінійної регресії для прогнозування рівня злочинності на душу населення.

****

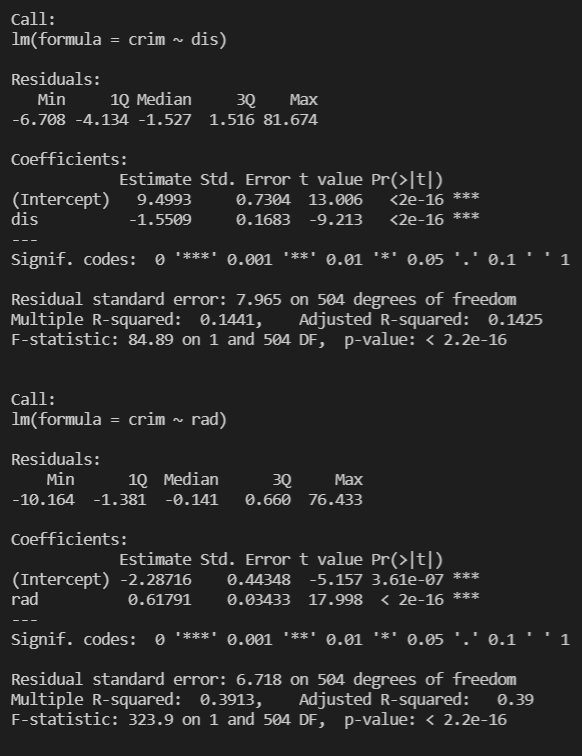
Модель лінійної регресії для для показника zn та indus

****

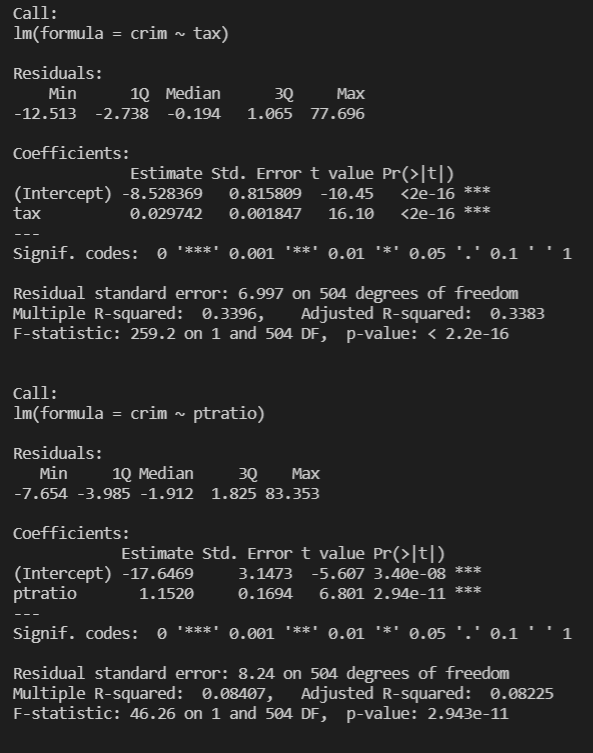
Модель лінійної регресії для для показника chas та nox

****

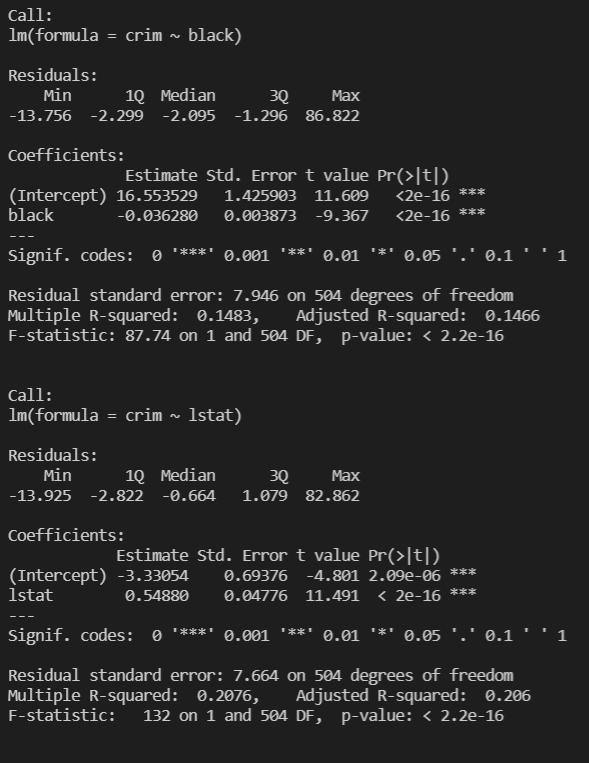
Модель лінійної регресії для для показника rm та age

****

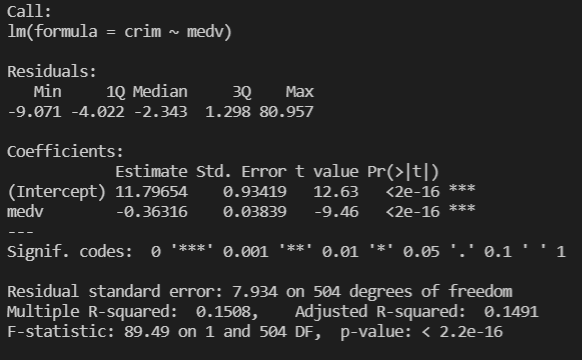
Модель лінійної регресії для для показника dis та rad

****

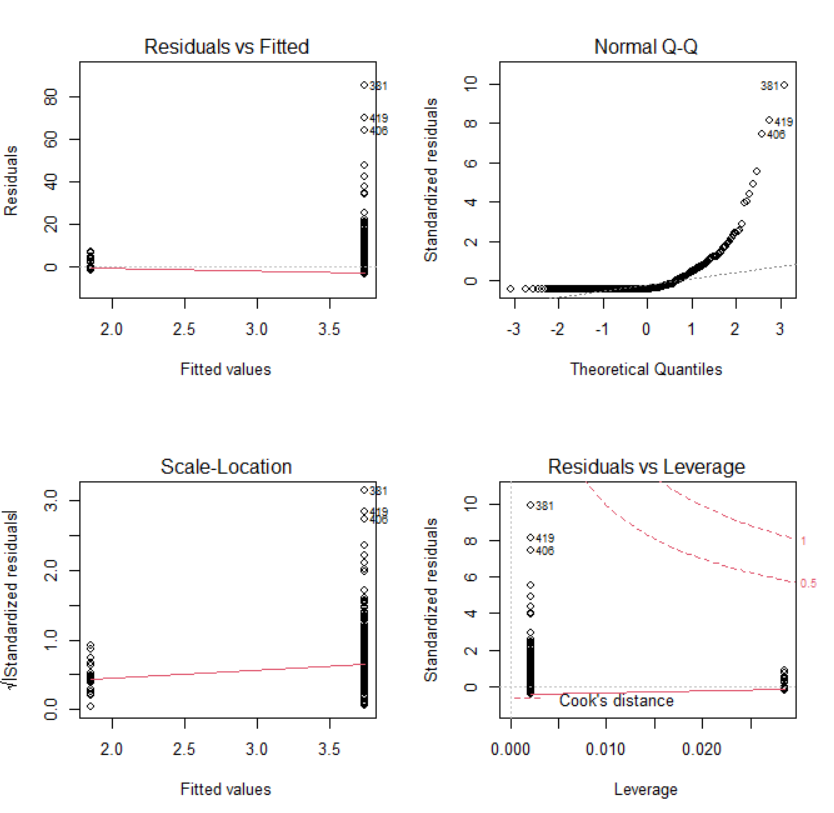
Модель лінійної регресії для для показника tax та ptratio

****

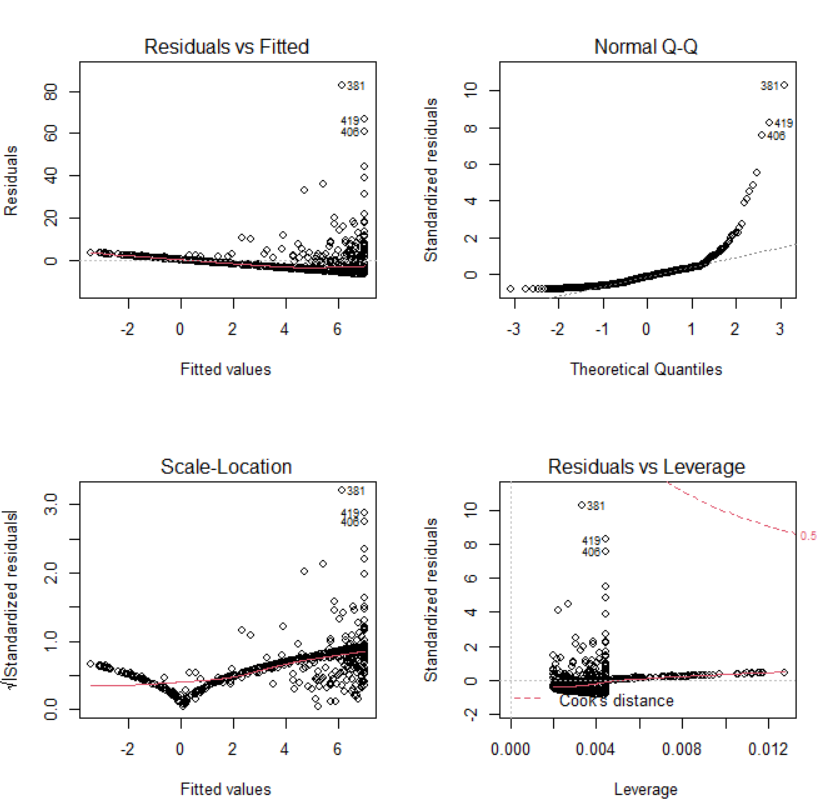
Модель лінійної регресії для для показника black та lstat

****

Модель лінійної регресії для для показника medv

****

Графік оцінки моделі із предиктором chas

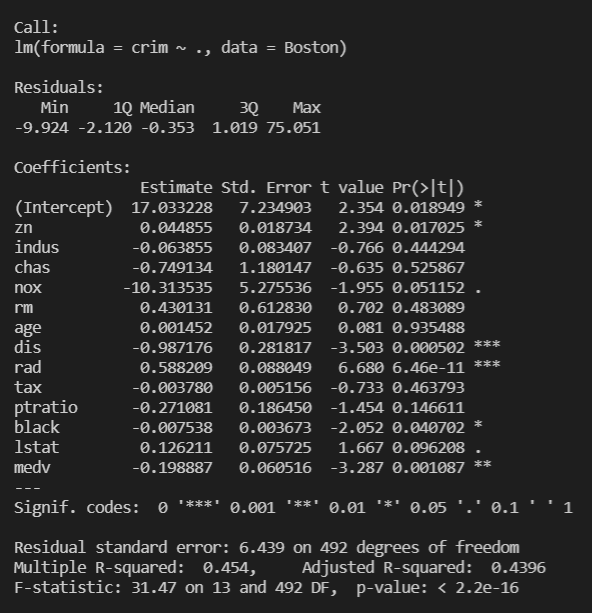
****

Графік оцінки моделі із предиктором age

Усі предиктори мають p-value менше 0,05, крім chas, тому ми можемо зробити висновок, що існує статистично значущий зв'язок між кожним предиктором та залежною змінною, за винятком предиктора chas.

**8.2**

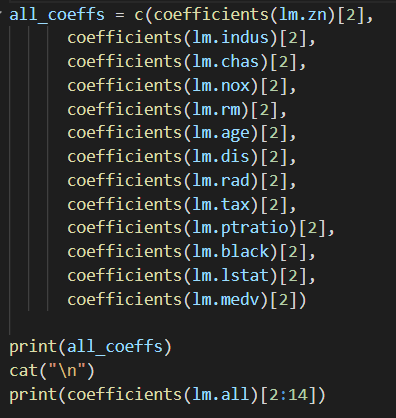
Побудовано модель множинної регресії для прогнозування залежної змінної за допомогою всіх предикторів.

****

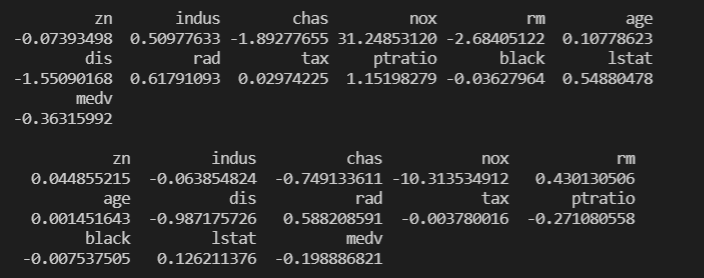
Ми можемо відхилити нульову гіпотезу для предикторів zn, dis, rad, black та medv.

**8.3**

Для порівняльного аналізу вищезгаданих моделей я використав функцію coefficients для потрібних моделей.

****

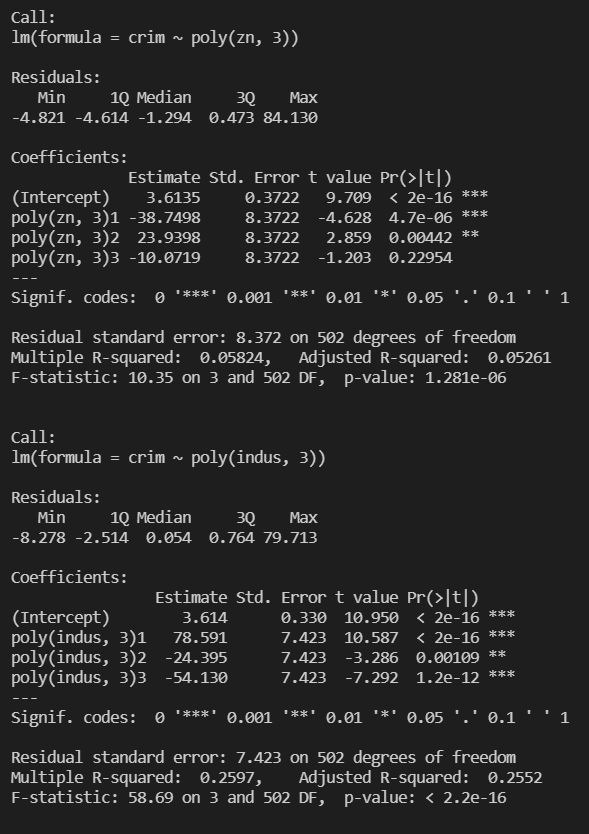
Можемо бачити нижче результати оціночного значення предикторів (перша таблиця – лінійна регресійна модель, друга – множинна регресійна модель).

****

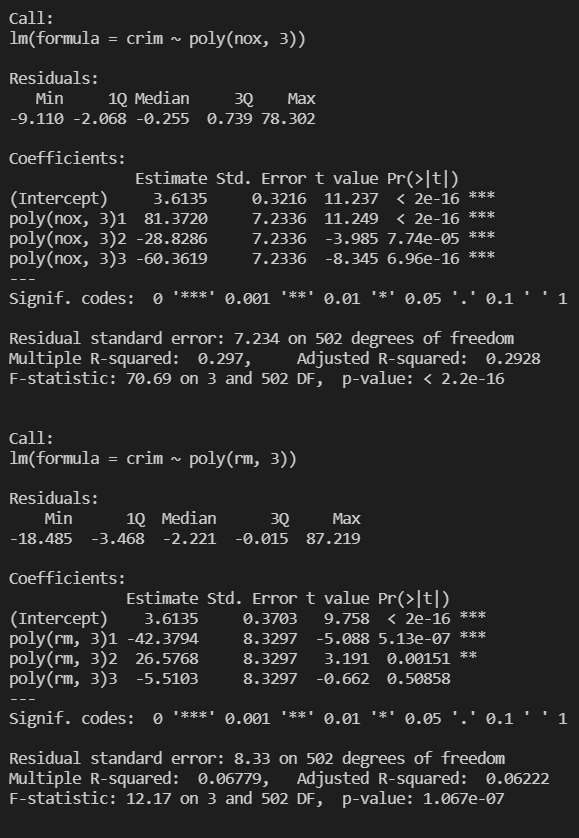
**8.4**

Для кожного предиктора серед даних Boston побудовано модель поліноміальної регресії до 3-го степеня завдяки функції poly(X, 3), де X це наш предиктор (рисунки наведені нижче).

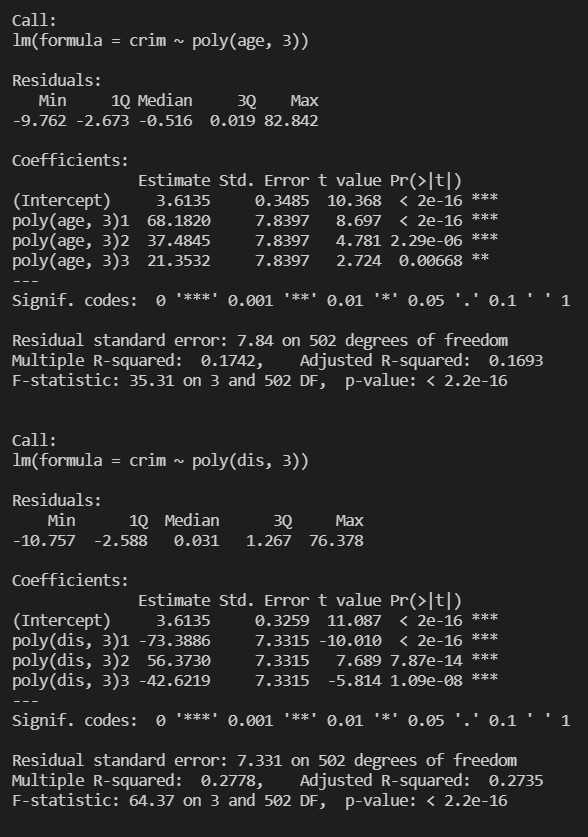
Оскільки змінна chas є якісною, то й для неї неможливо побудувати таку модель.



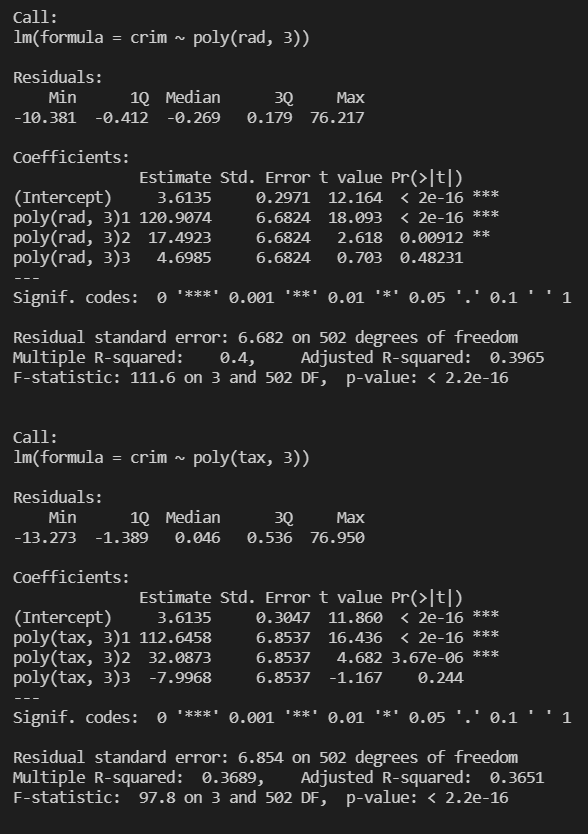
Модель поліноміальної регресії (3-го степеня) для показника zn та indus



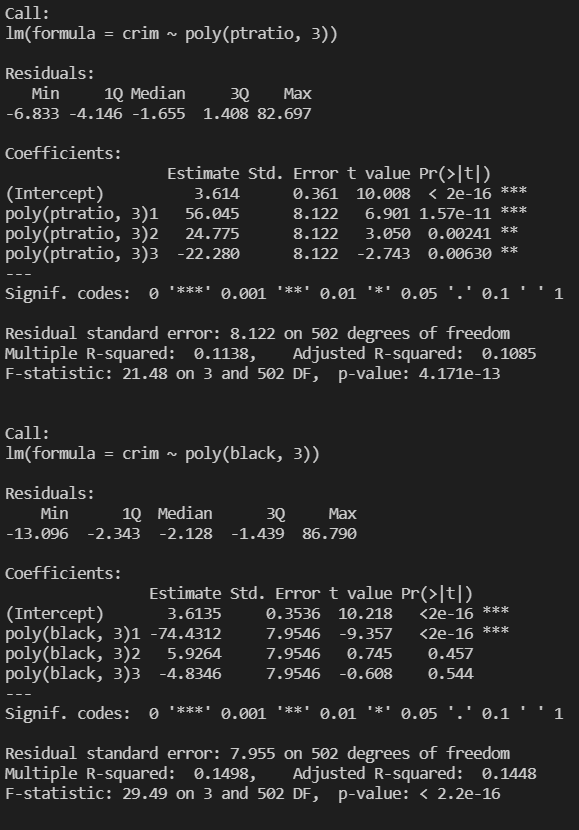
Модель поліноміальної регресії (3-го степеня) для показника nox та rm



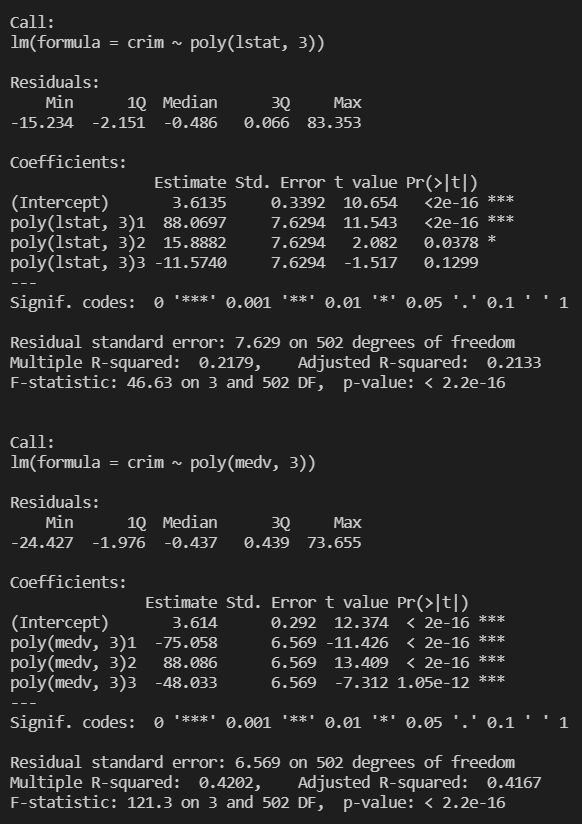
Модель поліноміальної регресії (3-го степеня) для показника age та dis



Модель поліноміальної регресії (3-го степеня) для показника rad та tax



Модель поліноміальної регресії (3-го степеня) для показника ptratio та black



Модель поліноміальної регресії (3-го степеня) для показника lstat та medv

Для предикторів zn, rm, rad, tax та lstat, p-values припускають, що кубічний коефіцієнт не є статистично значущим; для предиктора black, p-values припускають, що квадратичний та кубічний коефіцієнти не є статистично значущими, тому в цьому випадку нелінійного ефекту не видно.