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

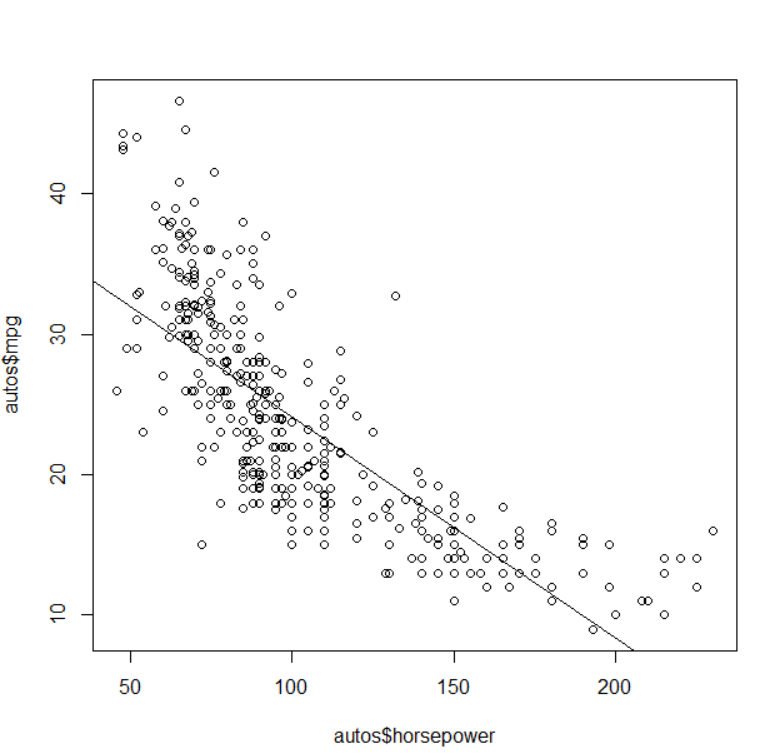
Так, існує залежність між horsepower та mpg, яка визначена шляхом перевірки нульової гіпотези всіх коефіцієнтів регресії, рівних нулю. Оскільки F-статистика набагато більша за 1, а p-значення F-статистики близьке до нуля, ми можемо відкинути нульову гіпотезу і стверджувати, що існує статистично значуща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

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

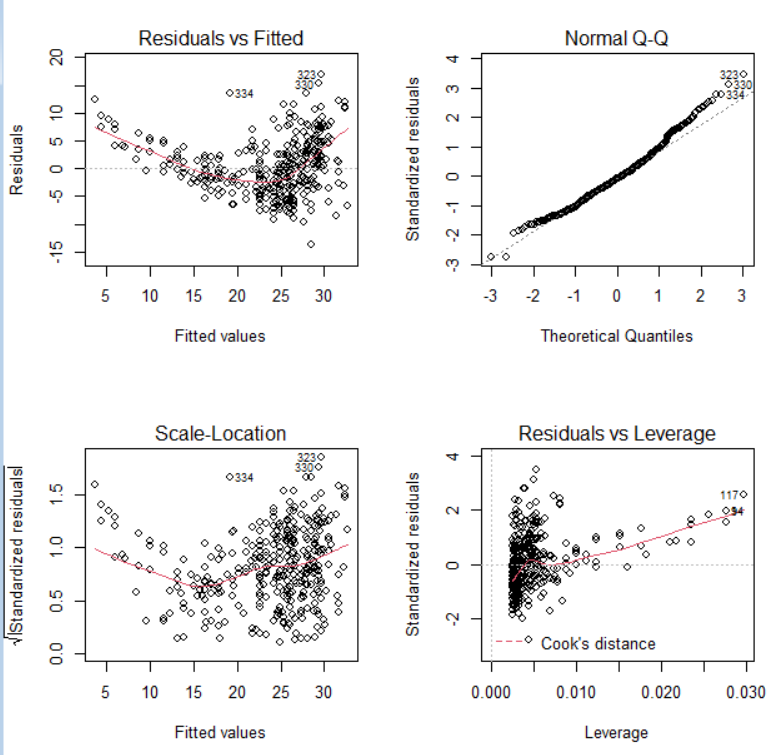
Зображення, що містить текст

Автоматично згенерований опис

**1.2**



**1.3**

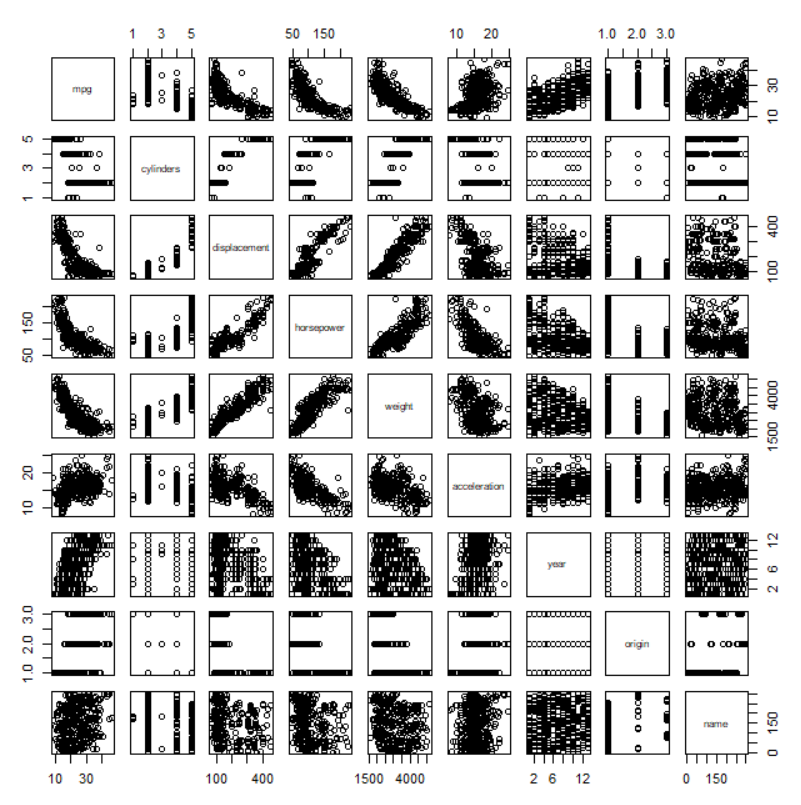


Зважаючи на ці графіки можна сказати, що залежність не зовсім лінійна.

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

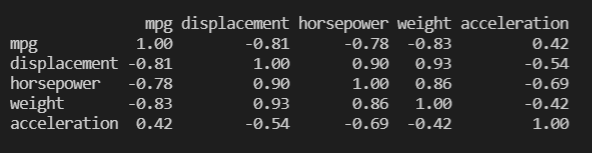
**2.1**

Побудовано діаграми розкиду усіх змінних.

****

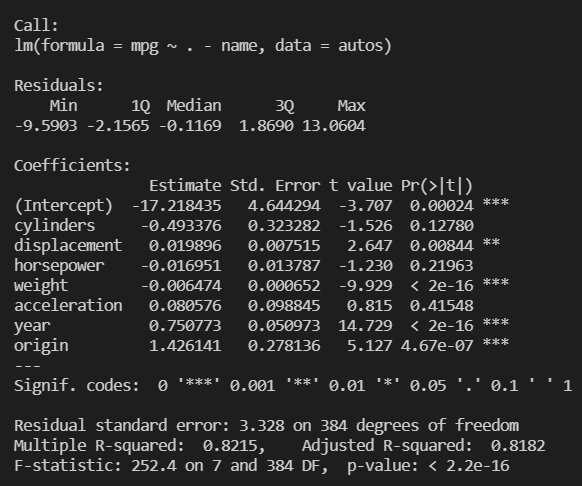
**2.2**

Обчислено матрицю кореляцій між змінними використовуючи функцію cor().

****

**2.3**

Використовуючи функцію lm() побудовано множинну регресію для залежної змінної mpg і всіх решту змінних окрім names як предикторів.

****

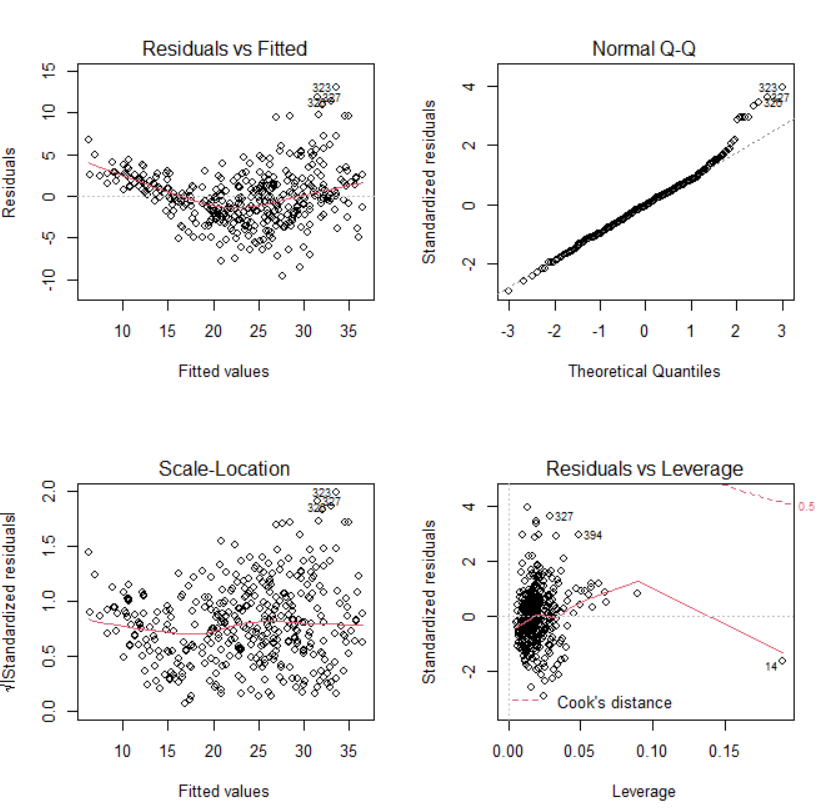
F-statistic є досить великою, тобто набагато більша за 1 з малим p-value, що свідчить проти нульової гіпотези про те що всі коефіцієнти регресії є нульовими, тобто є зв’язок між залежною змінною та предикторами.

З огляду на p-values, бачимо що displacement, weight, year та origin мають статистично значущий зв’язок із залежною змінною, тоді як cylinders, horsepower та acceleration ні.

Коефіцієнт регресії для year 0.75 свідчить про зростання mpg майже кожного року, що відбувається майже у відношенні 1 mpg/year.

**2.4**

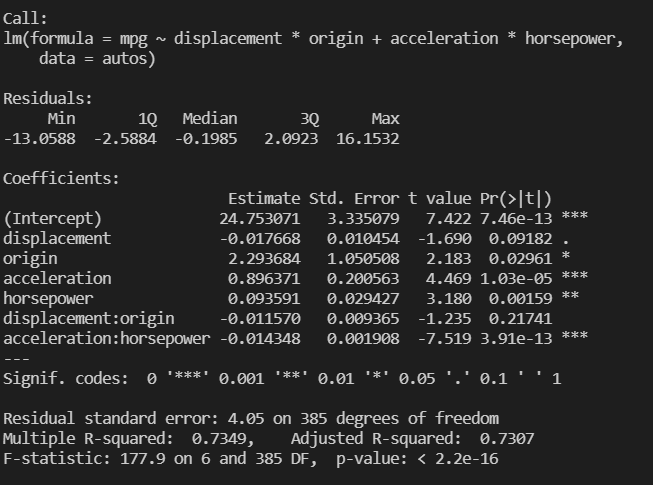
Використовуючи функцію plot () створено діагностичні графіки.

****

Зразу можна побачити, що модель є не дуже точною, оскільки на графіку Residuals vs Fitted є помітна крива, що свідчить про відхилення залишків. З графіку Residuals vs Leverage бачимо, що не є значно великими відхилення залишків і є точка (14) з високим левереджем.

**2.5**

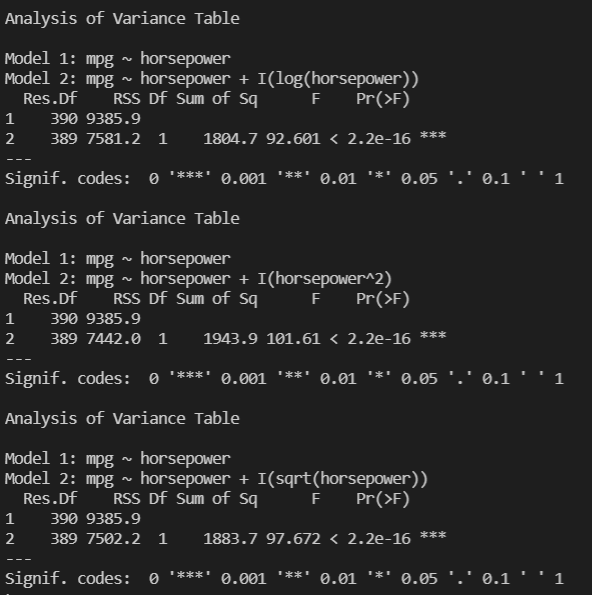
Використовуючи символ \* включив в модель лінійної регресії ефект взаємодії.

****

З p-values ми бачимо, що взаємодія між acceleration та horsepower є статистично значущою, тоді як взаємодія між displacement та origin не є такою.

**2.6**

Використано різні перетворення змінних. Як залежну змінну було взято mpg, а як предиктор horsepower. Порівняв базову модель з кожною з додаванням змінних, таких як log(X), *X*2, https://lh5.googleusercontent.com/wYueLr02gcDRmO0TZHseaFkv0-uOygtqMKIfb6fncvrTKRH_XQOfWLBazSo--dtUw_E9Zb7ohU_RLJHqZUALrbQdxMTXi2l4xii4KhrqLeCtyP0Z7KVF4zQd1zT1BaZmaWIanWg=s0. Використав для цього функцію anova().

****

З огляду на наведені вище таблиці можна впевнено сказати, що ми відхиляємо нульову гіпотезу про те що моделі з перетворенням змінних однаково добре описують дані, тобто повна модель в кожному випадку з нелінійним перетворенням є кращою.

**3. Розглянемо дані Carseats.**

Можемо побачити дані та їх опис.

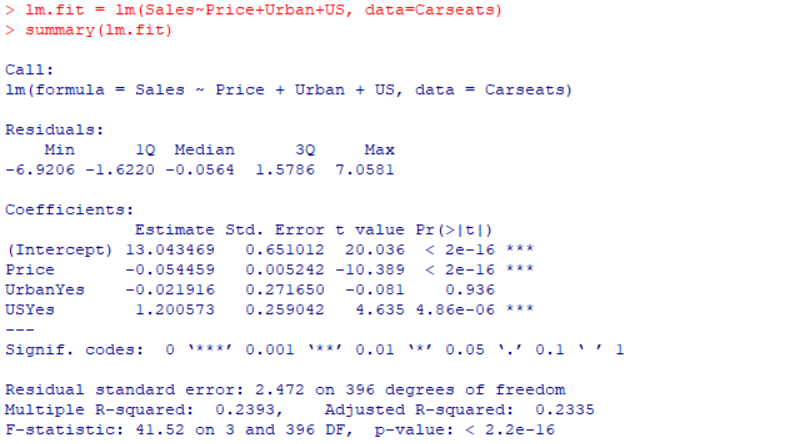
**Зображення, що містить стіл

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**

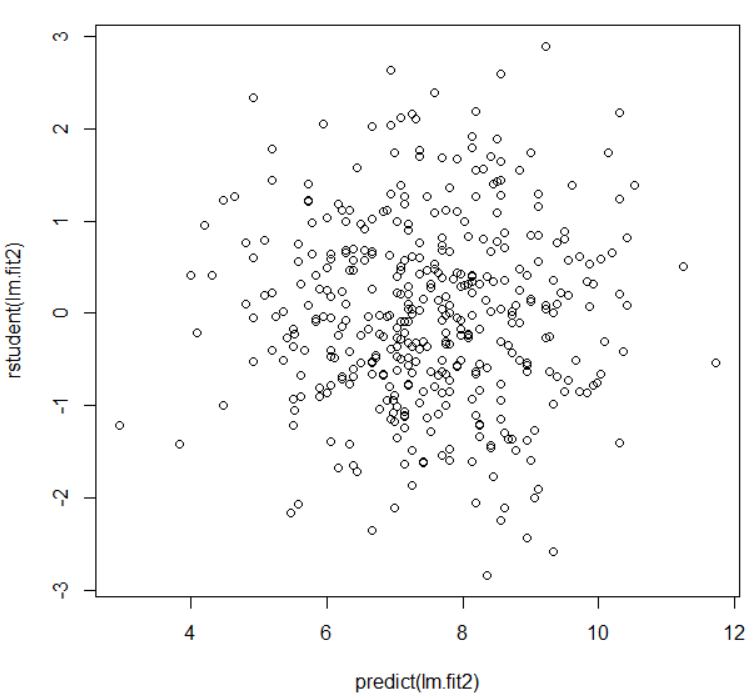
Зважаючи на значення RSE та можна стверджувати, що обидві моделі добре підходять для даних. Проте друга модель трошки краща.

**3.7**

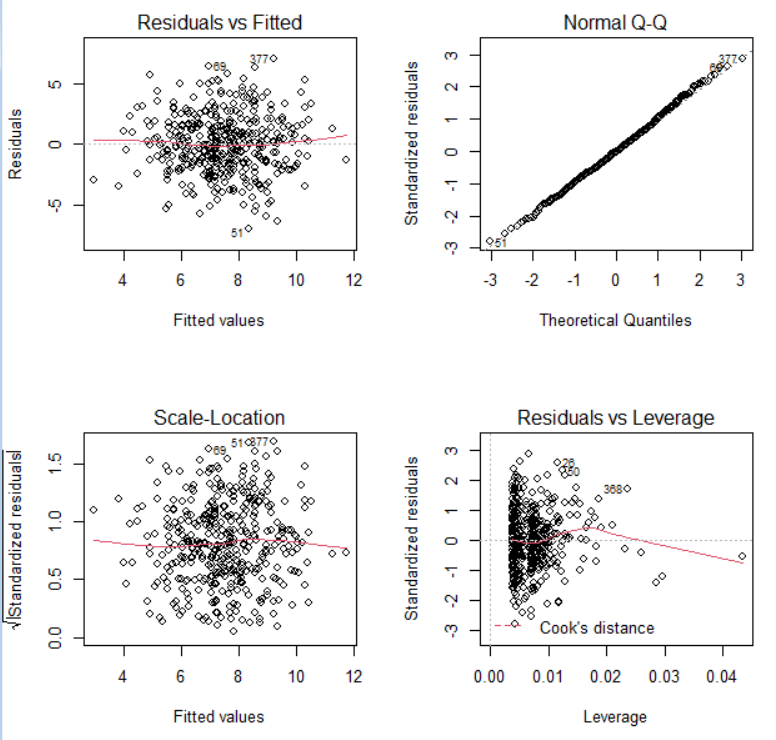
Зображення, що містить текст

Автоматично згенерований опис

**3.8**



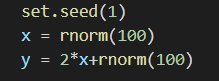
Усі Стюдентифіковані залишки, обмежені від -3 до 3, тому з лінійної регресії не випливають потенційні викиди.



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.

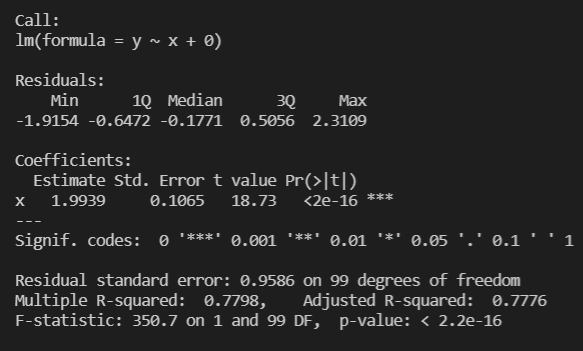
**4. Дослідження t-статистики для нульової гіпотези у простій лінійній регресії без коефіцієнта *β*0.**

Для початку ми згенеруємо предиктор *x* та залежну змінну *y*.



**4.1**

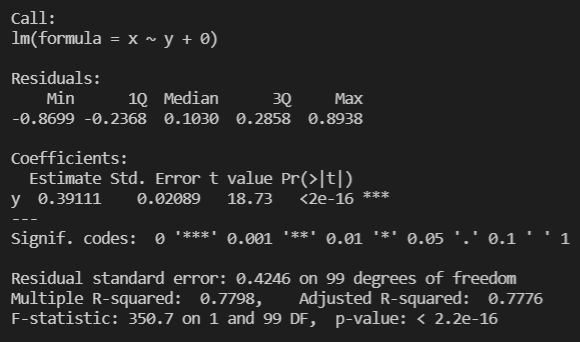
Побудовано просту лінійну регресію y на x без .

****

З наведеного вище p-value з t-статистики, яке є дуже малим (майже нульовим), можна зробити висновок про відкидання нульової гіпотези (H0: = 0).

**4.2**

Побудовано просту лінійну регресію x на y без .

****

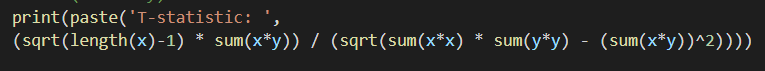
З наведеного вище p-value з t-статистики, бачимо аналогічний результат з попередньою моделлю, тобто ми відкидаємо нульову гіпотезу (H0: = 0).

**4.3**

****

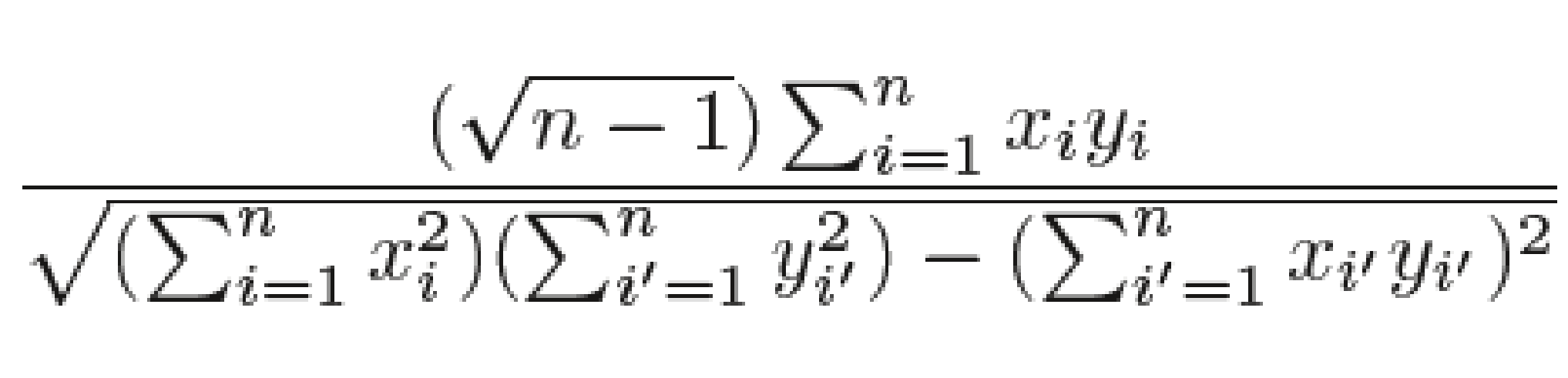
Можемо бачити досить тісну кореляцію змінних y та x. Про це свідчить і той факт, y=2x+ɛ може бути розписане через x=0.5∗(y− ɛ).

**4.4**

****

****

Чисельно перевірено, що справді t-статистика може бути записана в такому вигляді.



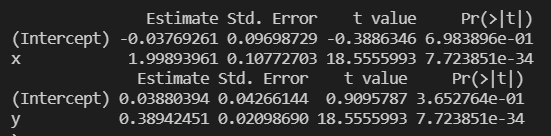
**4.5**

З огляду на наведені вище підсумкові дані по моделям лінійної регресії x на y та y на x бачимо що t-статистика для обох є однаковою t value=18.73.

Також з огляду на формулу t-статистики у пункті 4.4, то як бачимо значення не зміниться коли ми поміняємо місцями x та y (бо вони фігурують тільки в добутках).

**4.6**

Побудовано просту лінійну регресію з коефіцієнтом як для x на y, так й для y на x.

****

Бачимо, що як і з моделями без коефіцієнта t-статистика для обох моделей є однаковою з t value=18.56. При чому варто наголосити, що моделі з коефіцієнтом мають інше значення t value порівнюючи з попередніми моделями.

**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

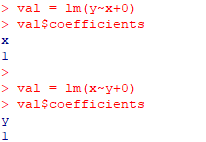
З результатів видно що вони різні

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

Зображення, що містить текст

Автоматично згенерований опис

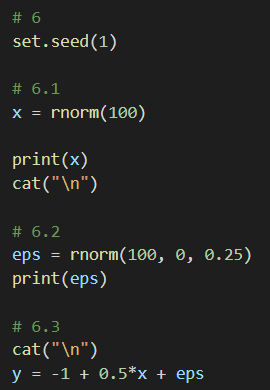
З результатів можна побачити, що коефіцієнти для лінійної регресії Y на X та X на Y однакові.



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

**6.1-6.3**

Створено вектор x та esp з використанням функції rnorm(). З них побудовано y відповідно до моделі y = – 1 + 0,5X + ε.



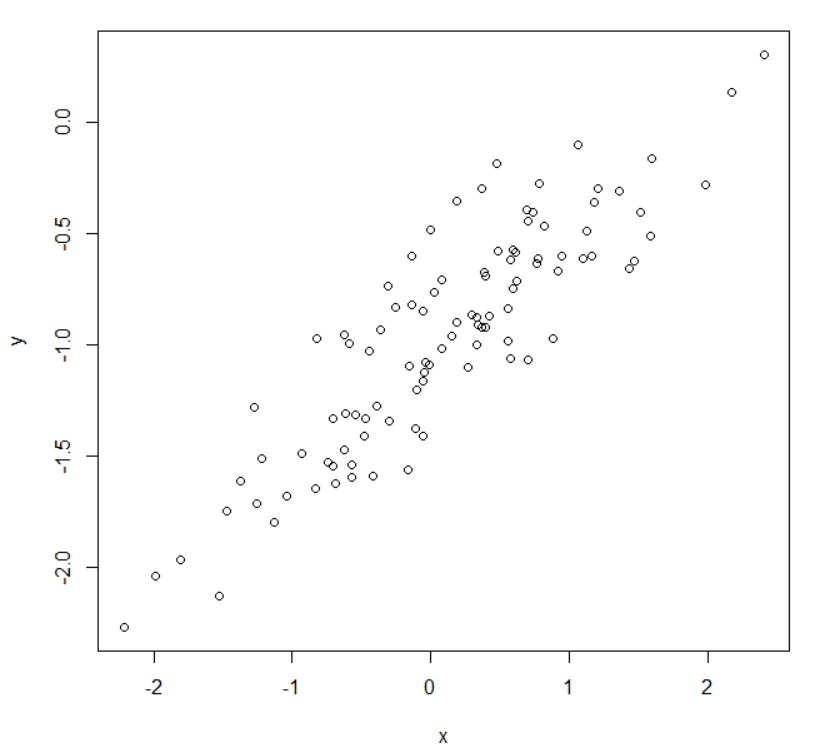
Як бачимо довжина вектора y - 100, =-1, а .

****

**6.4**

Побудовано діаграму розсіювання (рисунок нижче). Також показав досить тісну кореляцію між векторами x та y завдяки функції cor().

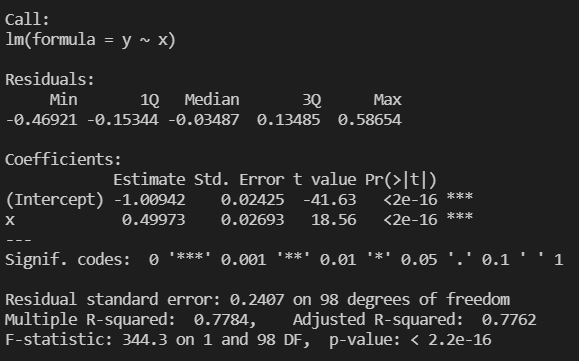
****

****

Відповідно до діаграми бачимо, що є досить чітка лінійна залежність між x та y.

**6.5**

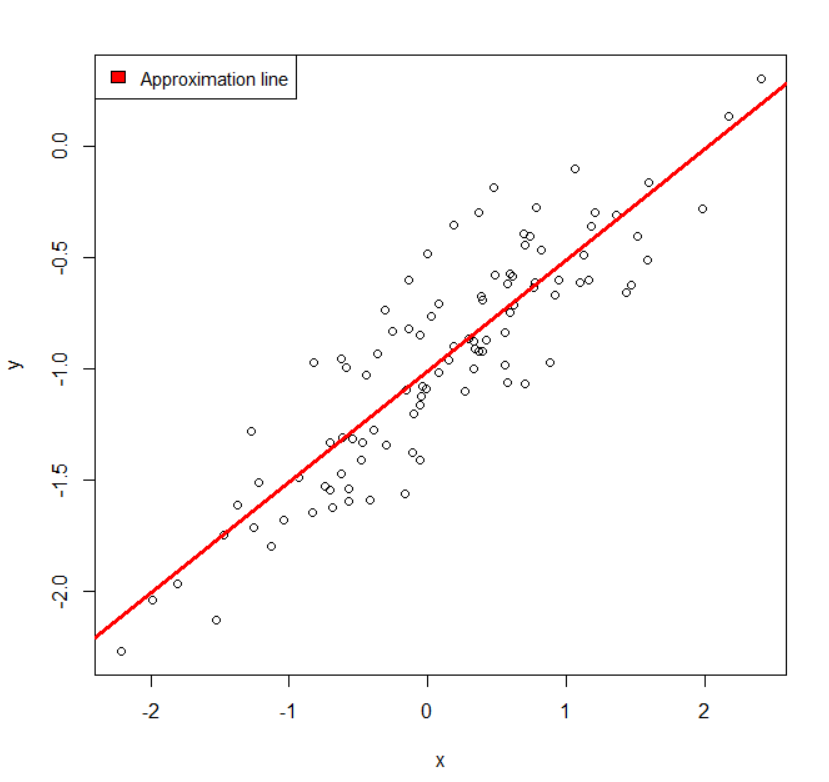
Побудовано лінійну модель для прогнозування y на основі x.

****

Взявши до уваги аналіз нашої моделі випливає, що значення параметрів β є дуже точними. А з огляду на низькі p-value з t-статистики, то наша лінійна модель є достовірною, що досить логічно взявши результати з пункту 6.5 про лінійну залежність x та y.

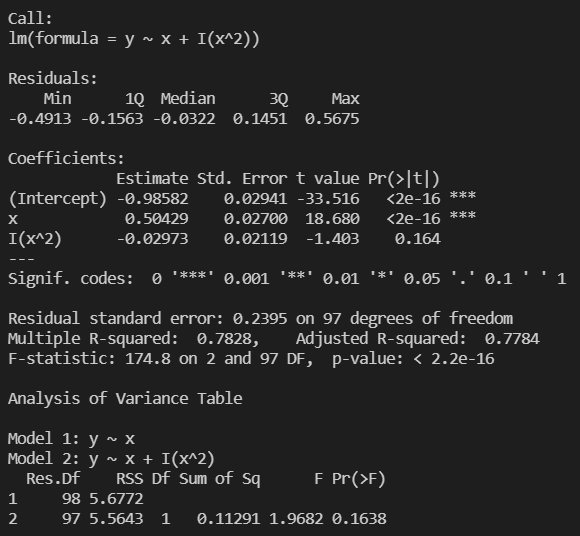
**6.6**

Побудовано оцінену лінію нашої моделі на діаграмі розсіяння.

****

**6.7**

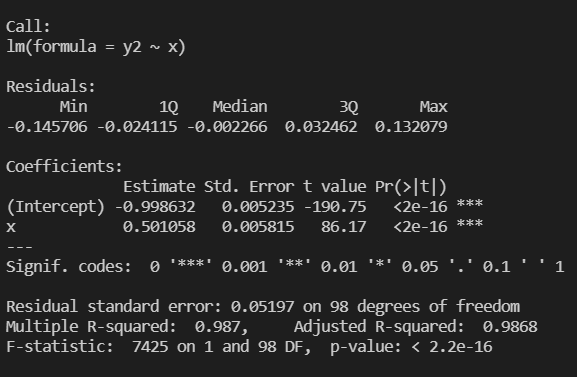
Побудовано модель поліноміальної регресії до 2-го степеня. Оцінка також здійснення з використанням функції anova().

****

Бачимо, що значення стало менш точним і p-value з t-статистики для квадратного коефіцієнта є досить великим, що тільки підтверджує лінійність x та y та робить цю модель менш придатною для наших даних.

**6.8**

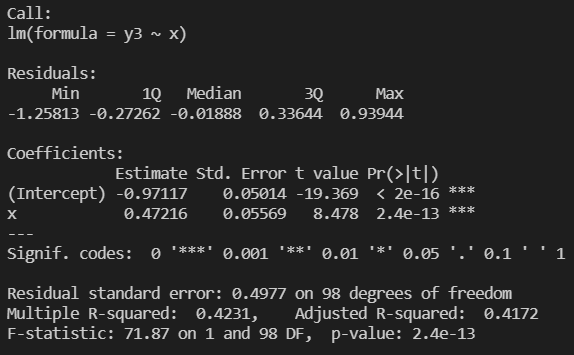
Повторено кроки 6.1-6.6 з модифікацією таким чином, щоб було менше шуму в даних (зменшено дисперсію для вектора ɛ до 0.05).

****

У підсумку можна сказати, що Multiple R-squared та Adjusted R-squared є дуже великими і майже повністю відповідають реальній регресії (98% відповідності).

**6.9**

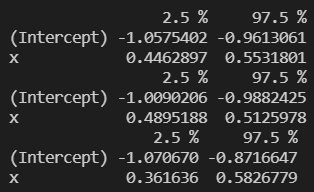
Повторено кроки 6.1-6.6 з модифікацією таким чином, щоб було більше шуму в даних (збільшено дисперсію для вектора ɛ до 0.5).

****

Як бачимо, що Multiple R-squared та Adjusted R-squared є досить низькими що свідчить про збільшення похибки нашої лінійної моделі.

**6.10**

Виведено довірчі інтервали для *β*0 та *β*1 на основі оригінальних даних, даних з більшим шумом та даних з меншим шумом.

****

Очевидно, що з збільшенням шуму довірчі інтервали збільшуються і навпаки.

**7. Зосередимося на проблемі колінеарності.**

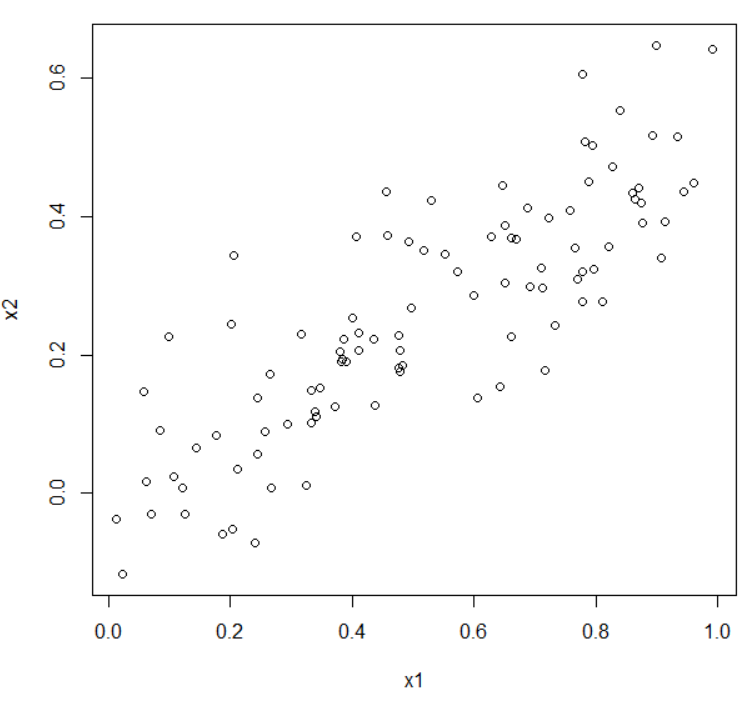
**7.1**

Форма лінійної моделі та коефіцієнти регресії

**7.2**

****

Кореляція між та та діаграма розсіювання



**7.3**

Зображення, що містить текст

Автоматично згенерований опис

Можемо побачити що коефіцієнт близький до реального значення і значення p надзвичайно мале що вказує на те, що модель підібрала його коректною.

Значення достатньо близьке до реальногоі його значення p може бути достатньо для відхилення нульової гіпотези.

Значення найгірше, а найвище значення p означає, що ми можемо прийняти 0 гіпотезу.

**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

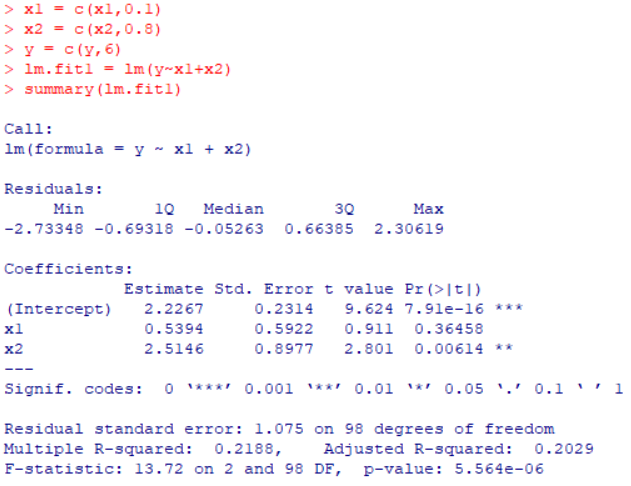
Автоматично згенерований опис

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**



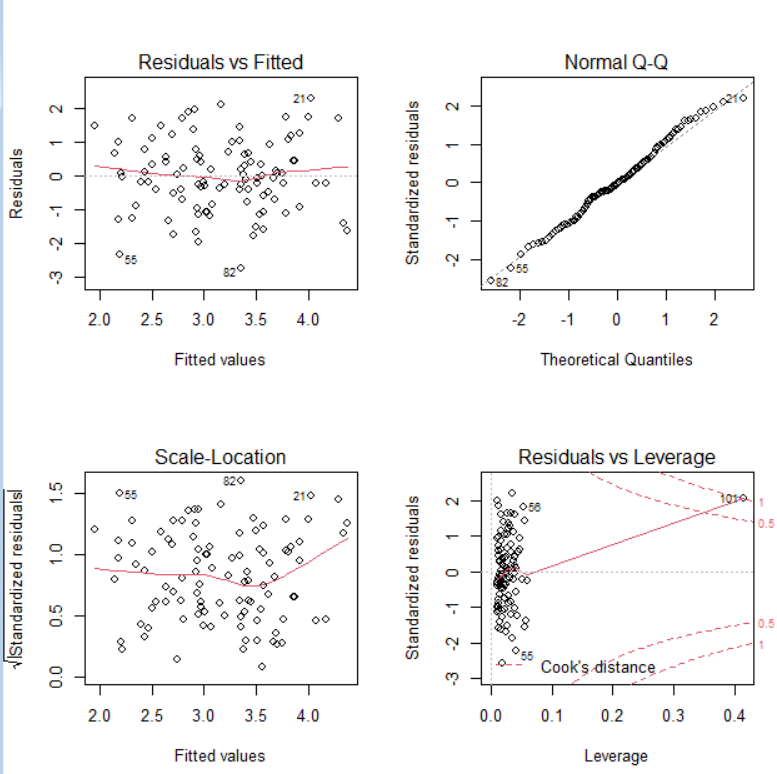
Зображення, що містить текст

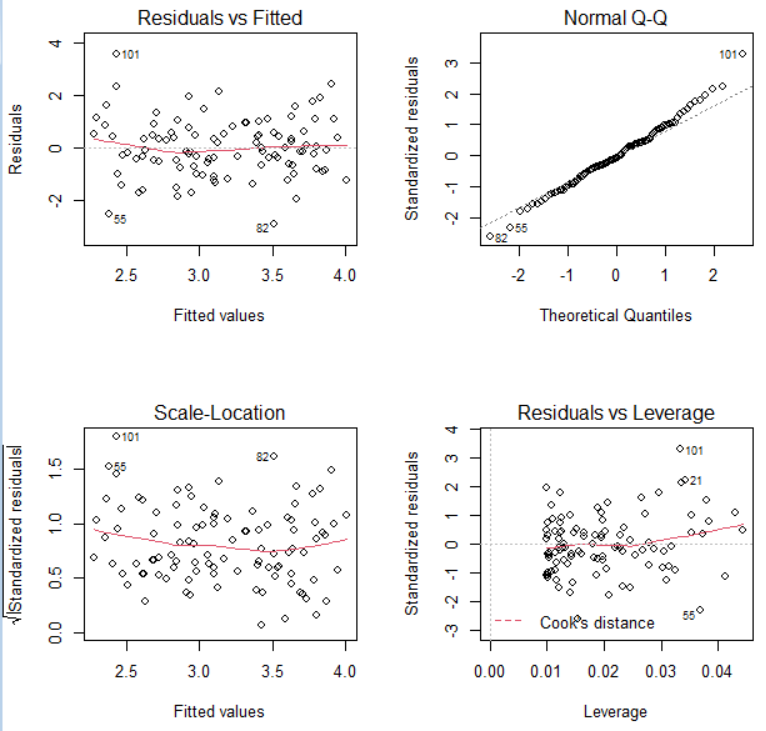
Автоматично згенерований опис

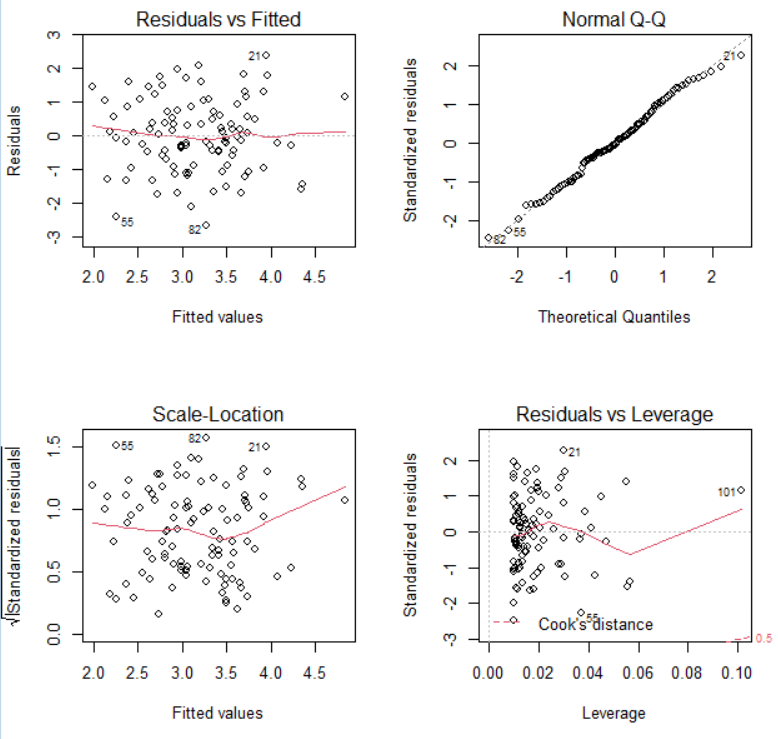
Зображення, що містить текст

Автоматично згенерований опис

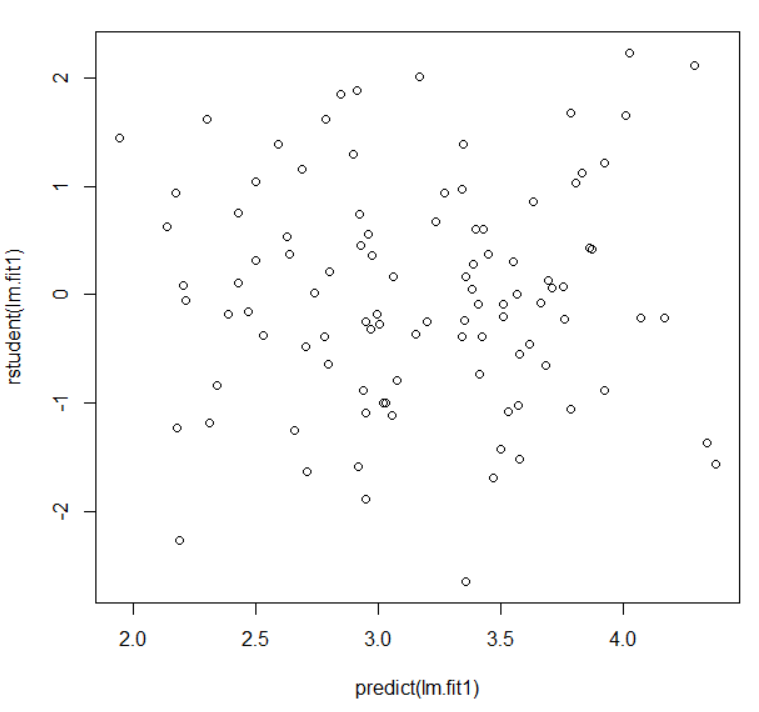
У першій моделі вона зсуває до статистично незначущої та до статистично значущої від зміни p-значень між двома лінійними регресіями.

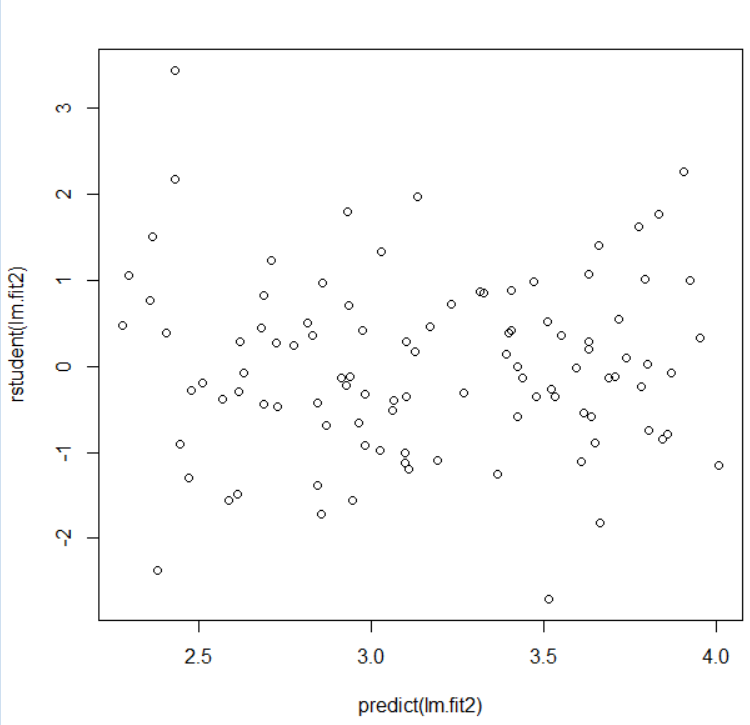


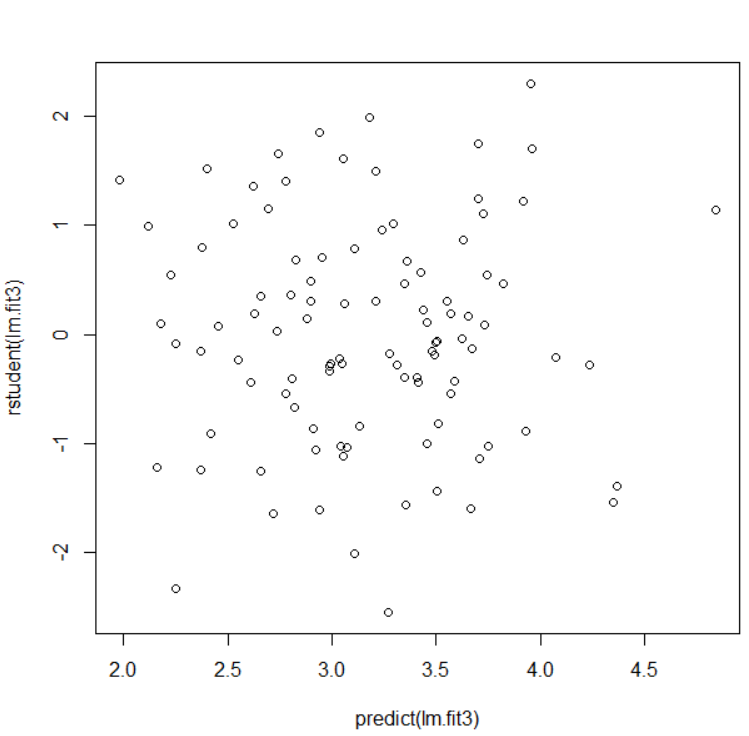




У першій і третій моделях ця точка стає точкою високого leverage.

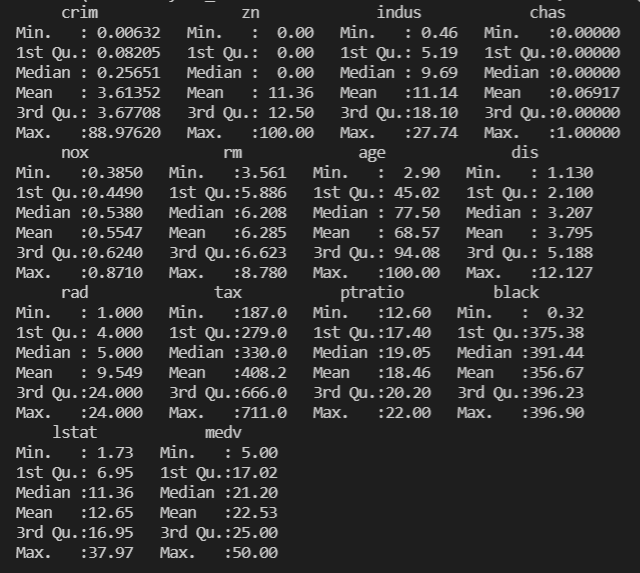






Дивлячись на Стюдентифіковані залишки, ми не спостерігаємо точок занадто далеко від граничного значення, що рівне , за винятком другої лінійної регресії: .

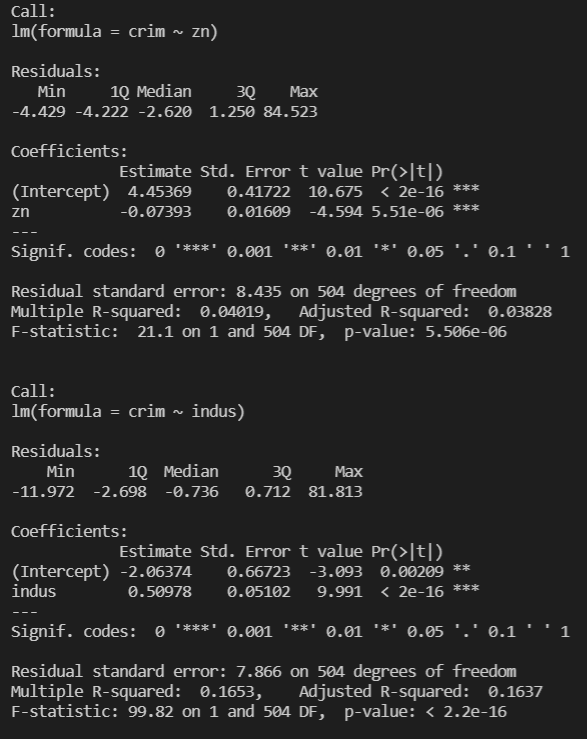
**8. Прогнозування рівня злочинності на душу населення використовуючи інші змінні в наборі даних Boston.**

****

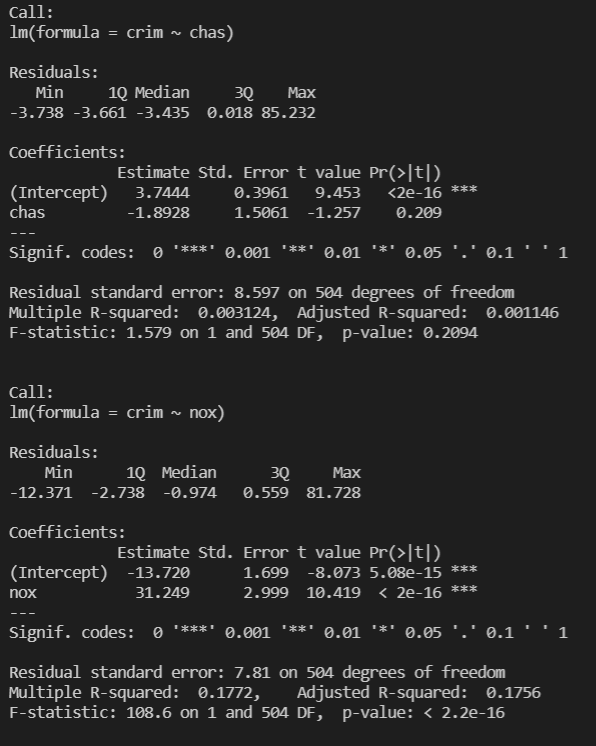
Загальна характеристика даних Boston

**8.1**

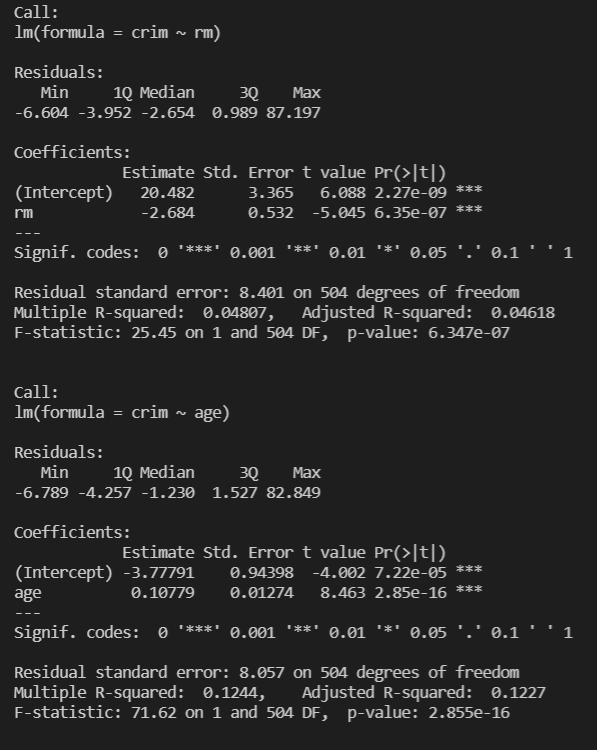
Побудовано для кожного предиктора просту модель лінійної регресії для прогнозування рівня злочинності на душу населення.

****

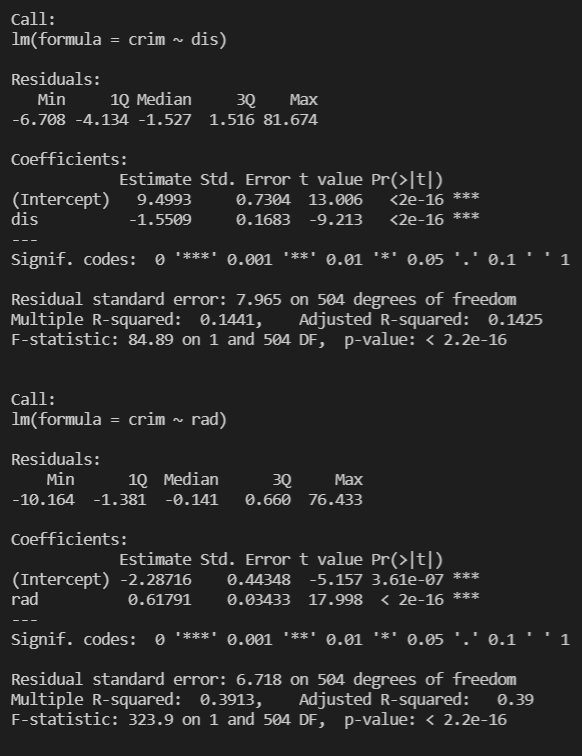
Модель лінійної регресії для для показника zn та indus

****

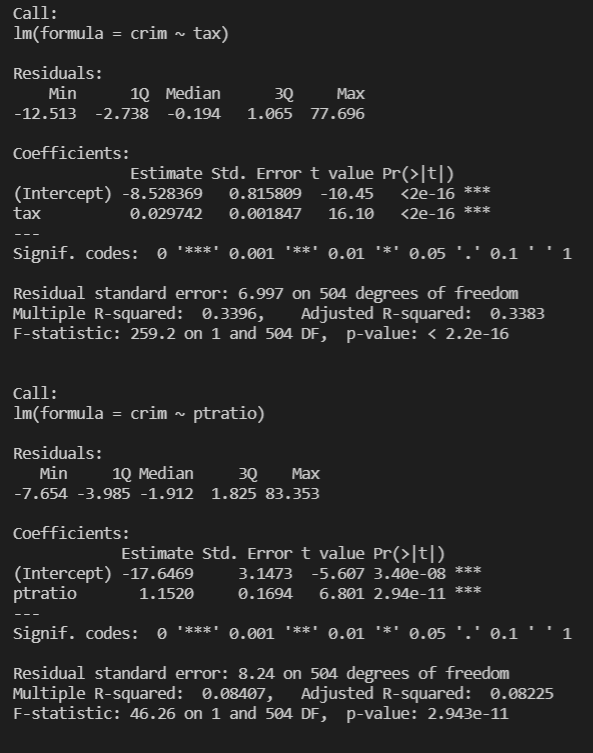
Модель лінійної регресії для для показника chas та nox

****

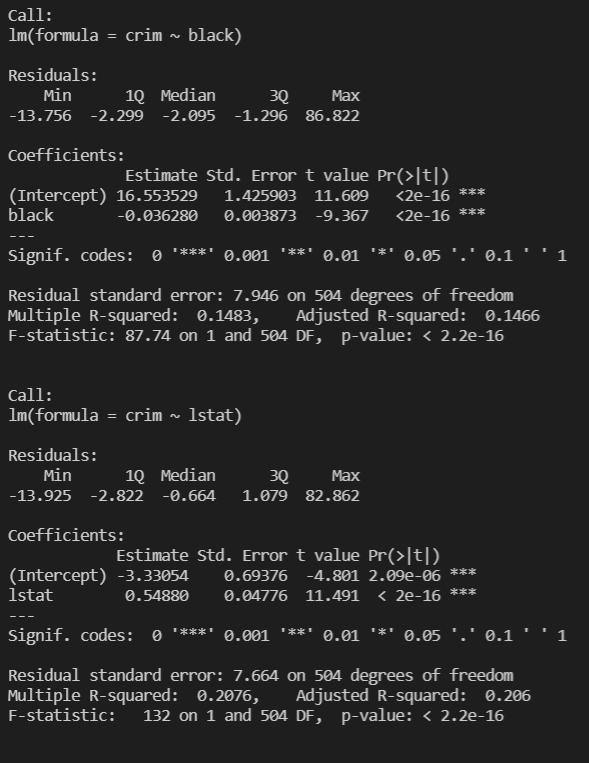
Модель лінійної регресії для для показника rm та age

****

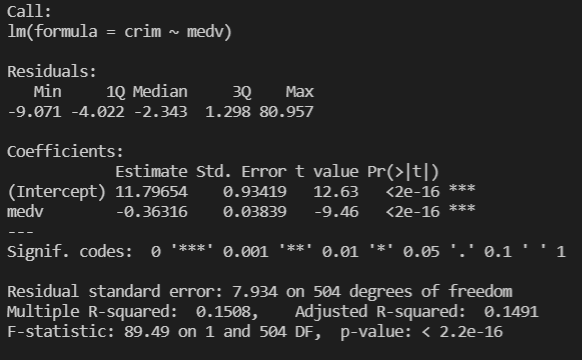
Модель лінійної регресії для для показника dis та rad

****

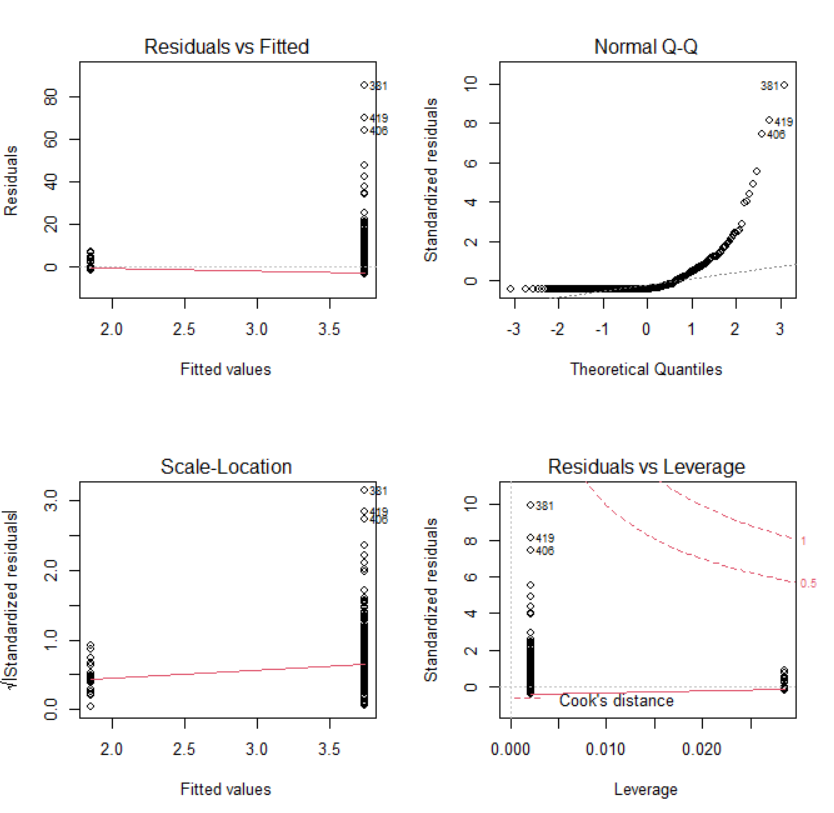
Модель лінійної регресії для для показника tax та ptratio

****

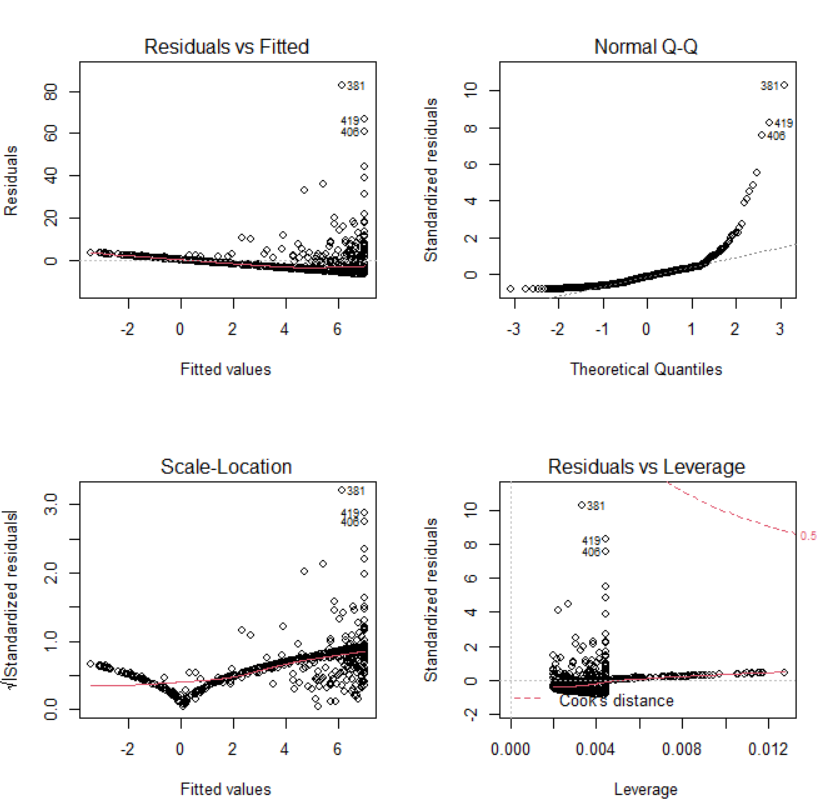
Модель лінійної регресії для для показника black та lstat

****

Модель лінійної регресії для для показника medv

****

Графік оцінки моделі із предиктором chas

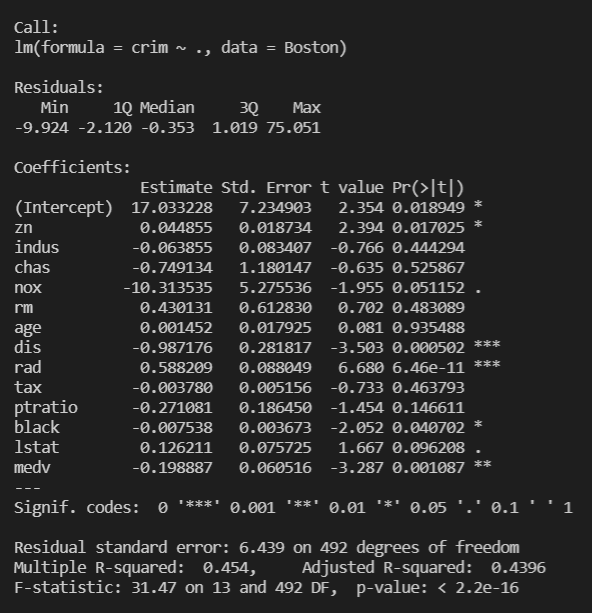
****

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**

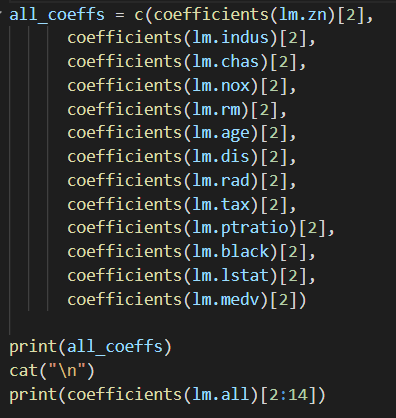
Побудовано модель множинної регресії для прогнозування залежної змінної за допомогою всіх предикторів.

****

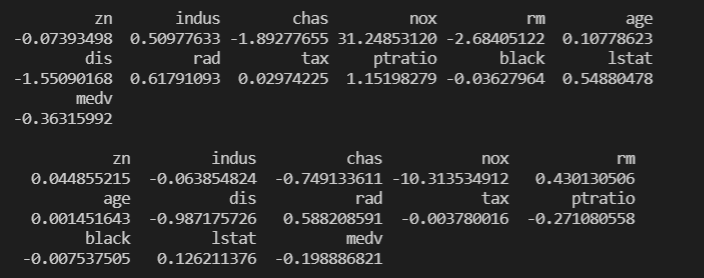
Ми можемо відхилити нульову гіпотезу для предикторів zn, dis, rad, black та medv.

**8.3**

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

****

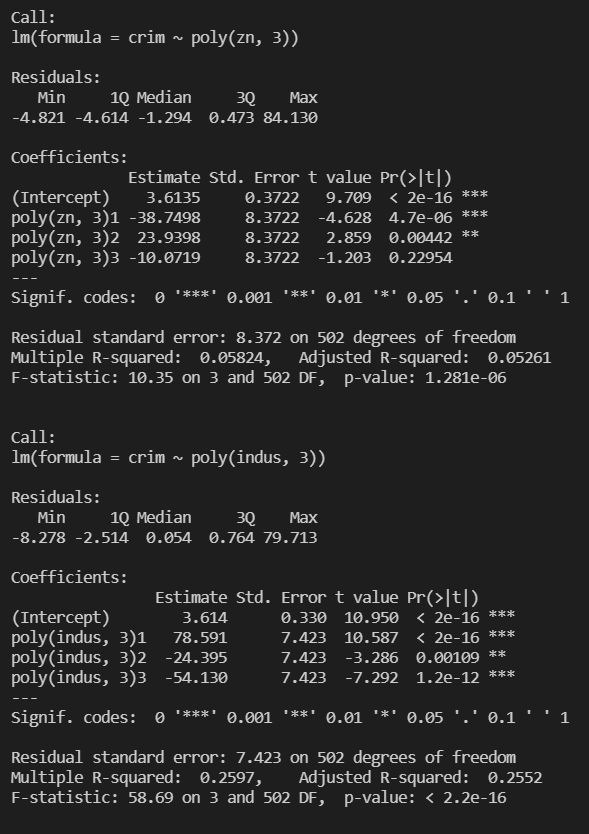
Можемо бачити нижче результати оціночного значення коефіцієнтів для предикторів (перша таблиця – лінійна регресійна модель, друга – множинна регресійна модель). Результати сильно відрізняються.

****

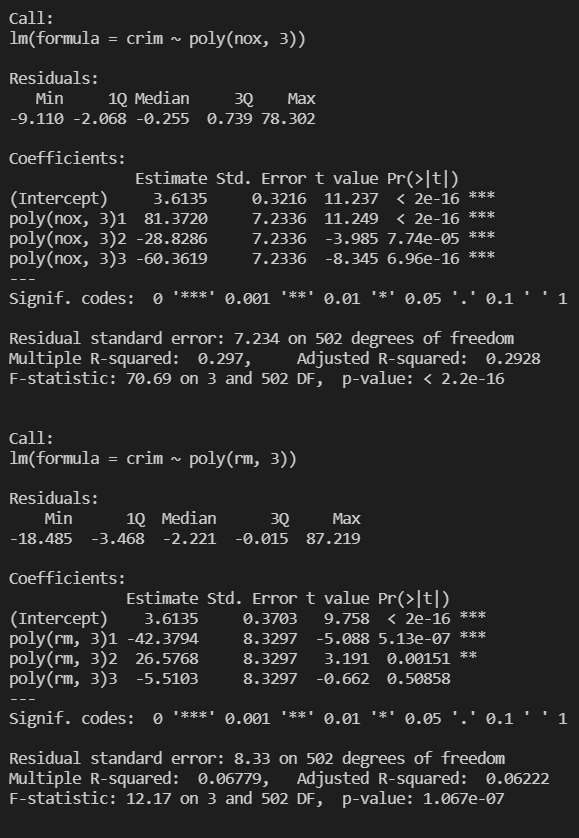
**8.4**

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).

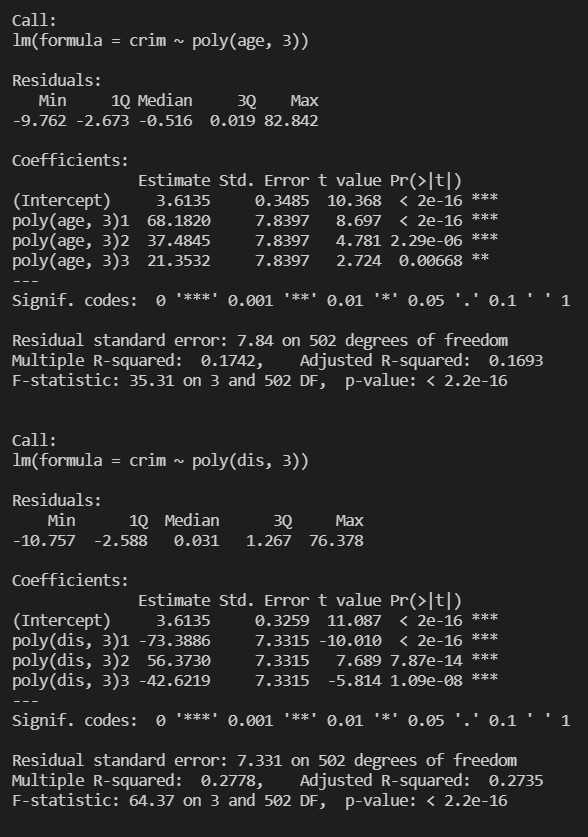
Оскільки змінна chas є якісною, то й для неї неможливо побудувати таку модель.



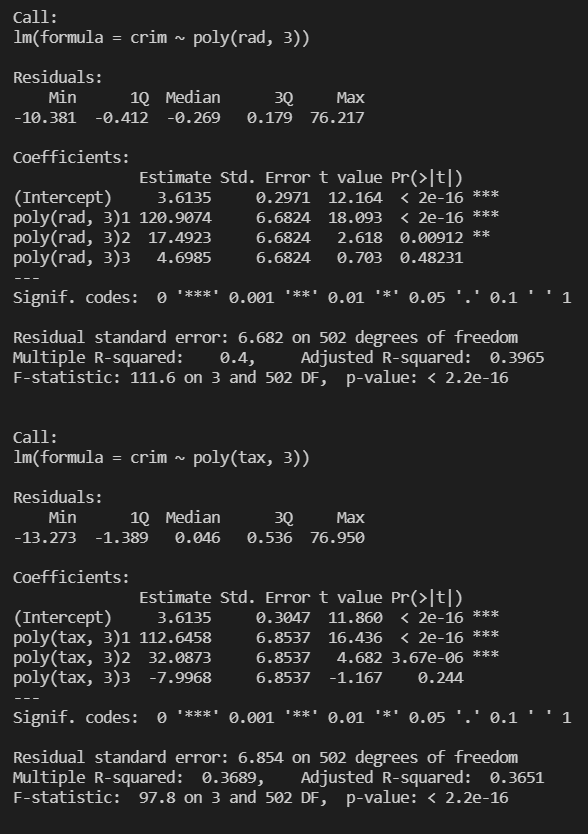
Модель поліноміальної регресії (3-го степеня) для показника zn та indus



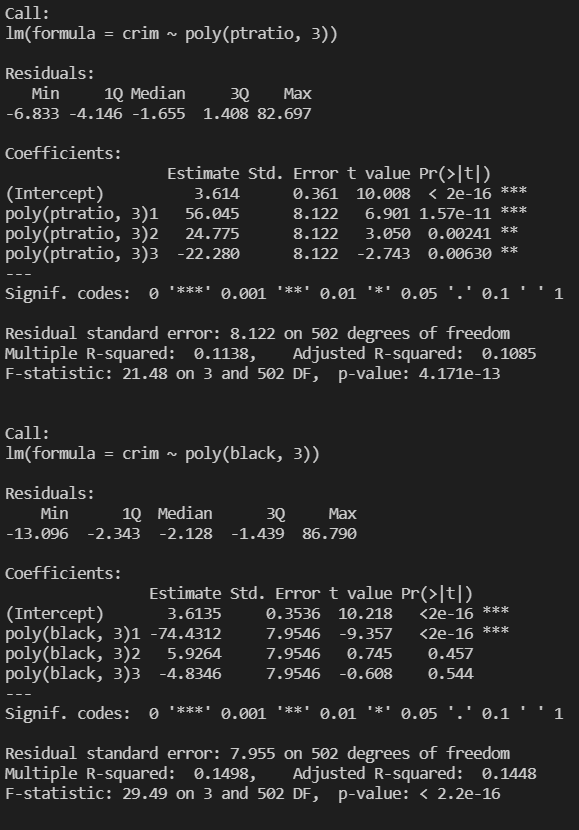
Модель поліноміальної регресії (3-го степеня) для показника nox та rm



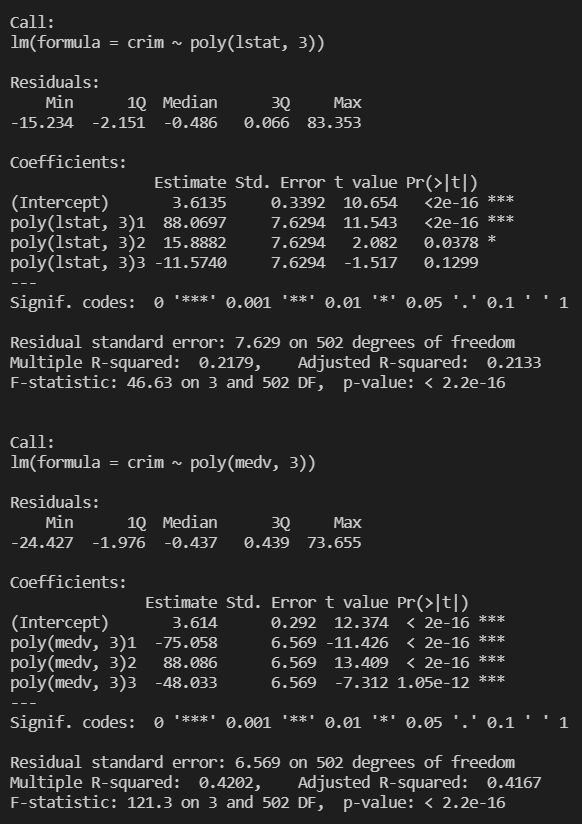
Модель поліноміальної регресії (3-го степеня) для показника age та dis



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.