

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

```
> lm.fit = lm(mpg~horsepower,data=autos)
> lm.fit

Call:
lm(formula = mpg ~ horsepower, data = autos)

Coefficients:
(Intercept)  horsepower
    39.9359      -0.1578

> summary(lm.fit)

Call:
lm(formula = mpg ~ horsepower, data = autos)

Residuals:
    Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  39.935861   0.717499   55.66  <2e-16 ***
horsepower   -0.157845   0.006446  -24.49  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

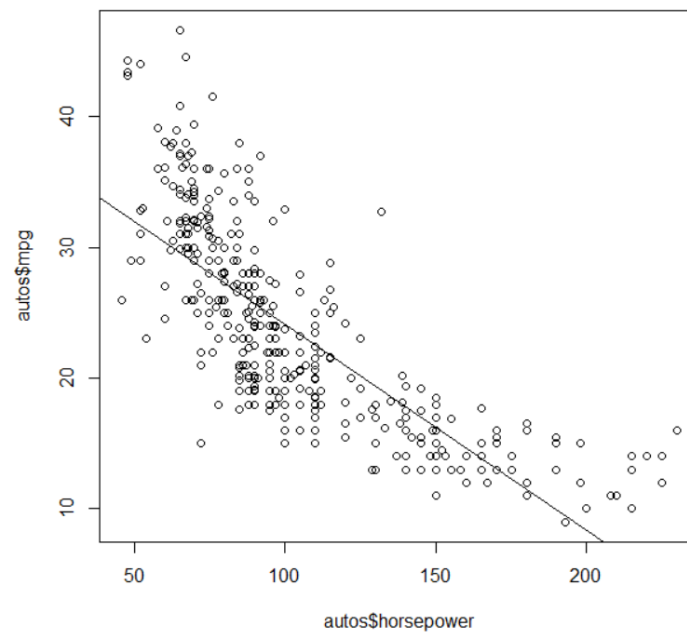
Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

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 р-значення F-статистики близьке до нуля, ми можемо відкинути нульову гіпотезу і стверджувати, що існує статистично значуща залежність між horsepower та mpg.

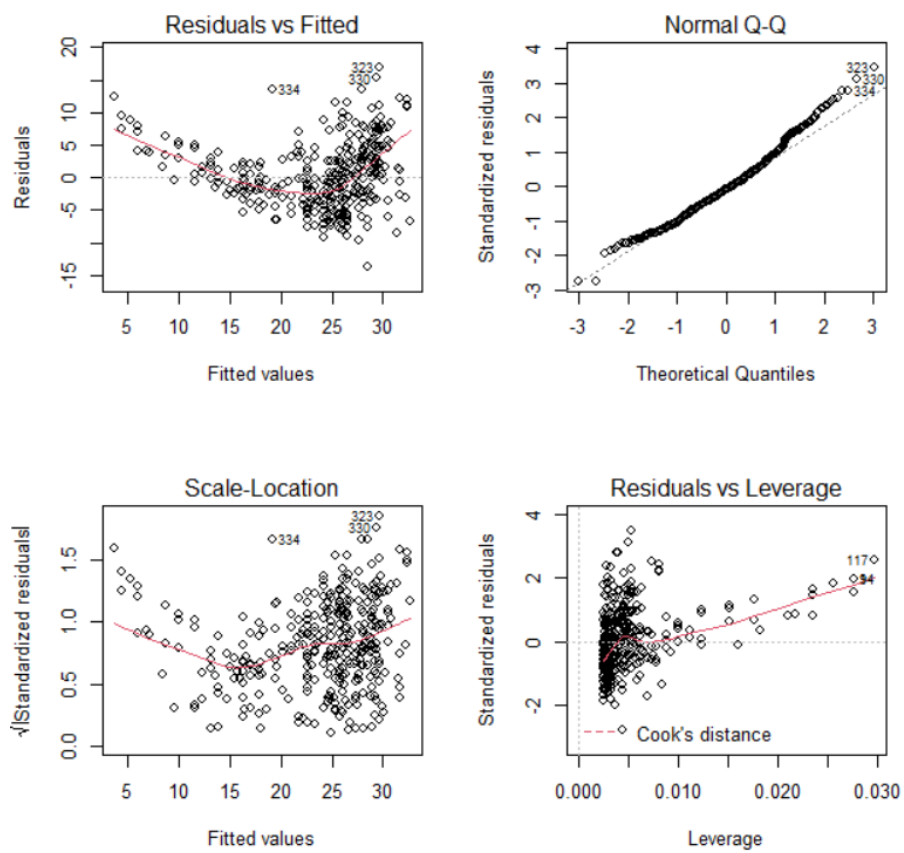
Співвідношення між mpg і horsepower є негативним. Чим більше horsepower в автомобіля, тим меншою є mpg автомобіля.

```
> predict(lm.fit, data.frame(horsepower=c(98)), interval="confidence")
      fit      lwr      upr
1 24.46708 23.97308 24.96108
> predict(lm.fit, data.frame(horsepower=c(98)), interval="prediction")
      fit      lwr      upr
1 24.46708 14.8094 34.12476
```

1.2



1.3

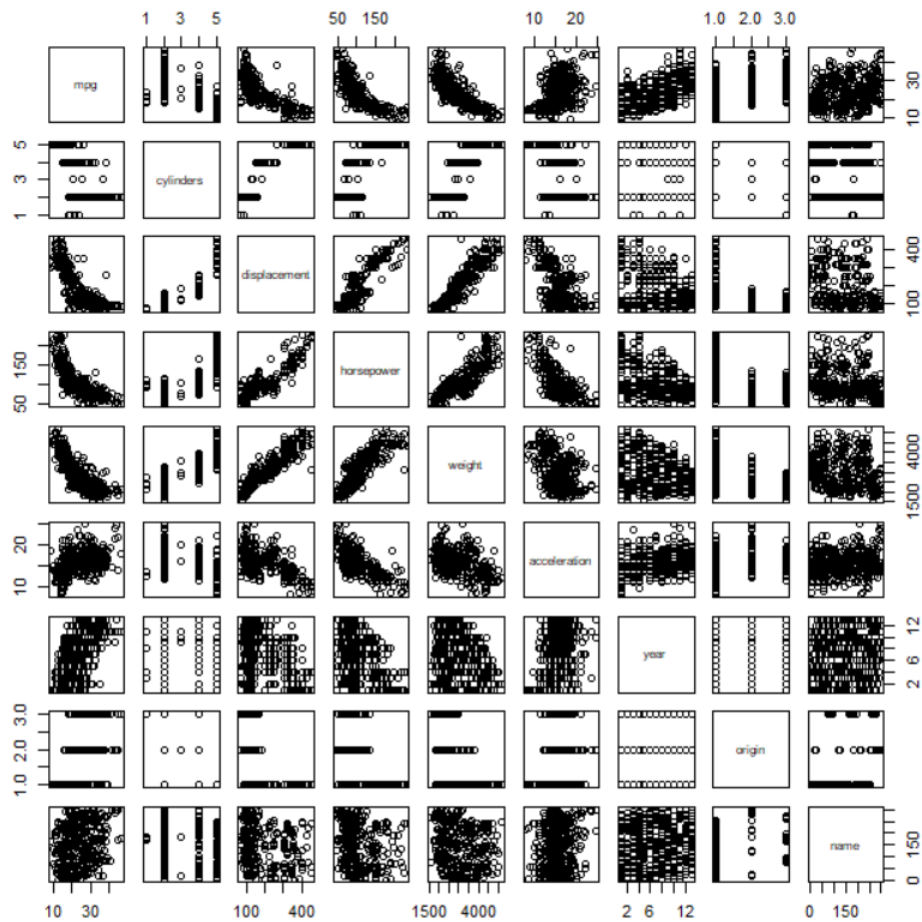


Зважаючи на ці графіки можна сказати, що залежність не зовсім лінійна.

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()`.

	mpg	displacement	horsepower	weight	acceleration
mpg	1.00	-0.81	-0.78	-0.83	0.42
displacement	-0.81	1.00	0.90	0.93	-0.54
horsepower	-0.78	0.90	1.00	0.86	-0.69
weight	-0.83	0.93	0.86	1.00	-0.42
acceleration	0.42	-0.54	-0.69	-0.42	1.00

2.3

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

```
Call:
lm(formula = mpg ~ . - name, data = autos)

Residuals:
    Min       1Q   Median       3Q      Max
-9.5903 -2.1565 -0.1169  1.8690 13.0604

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.21835    4.644294  -3.707  0.00024 ***
cylinders      -0.493376    0.323282  -1.526  0.12780
displacement   0.019896    0.007515   2.647  0.00844 **
horsepower    -0.016951    0.013787  -1.230  0.21963
weight        -0.006474    0.000652  -9.929 < 2e-16 ***
acceleration   0.080576    0.098845   0.815  0.41548
year           0.750773    0.050973  14.729 < 2e-16 ***
origin         1.426141    0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,    Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

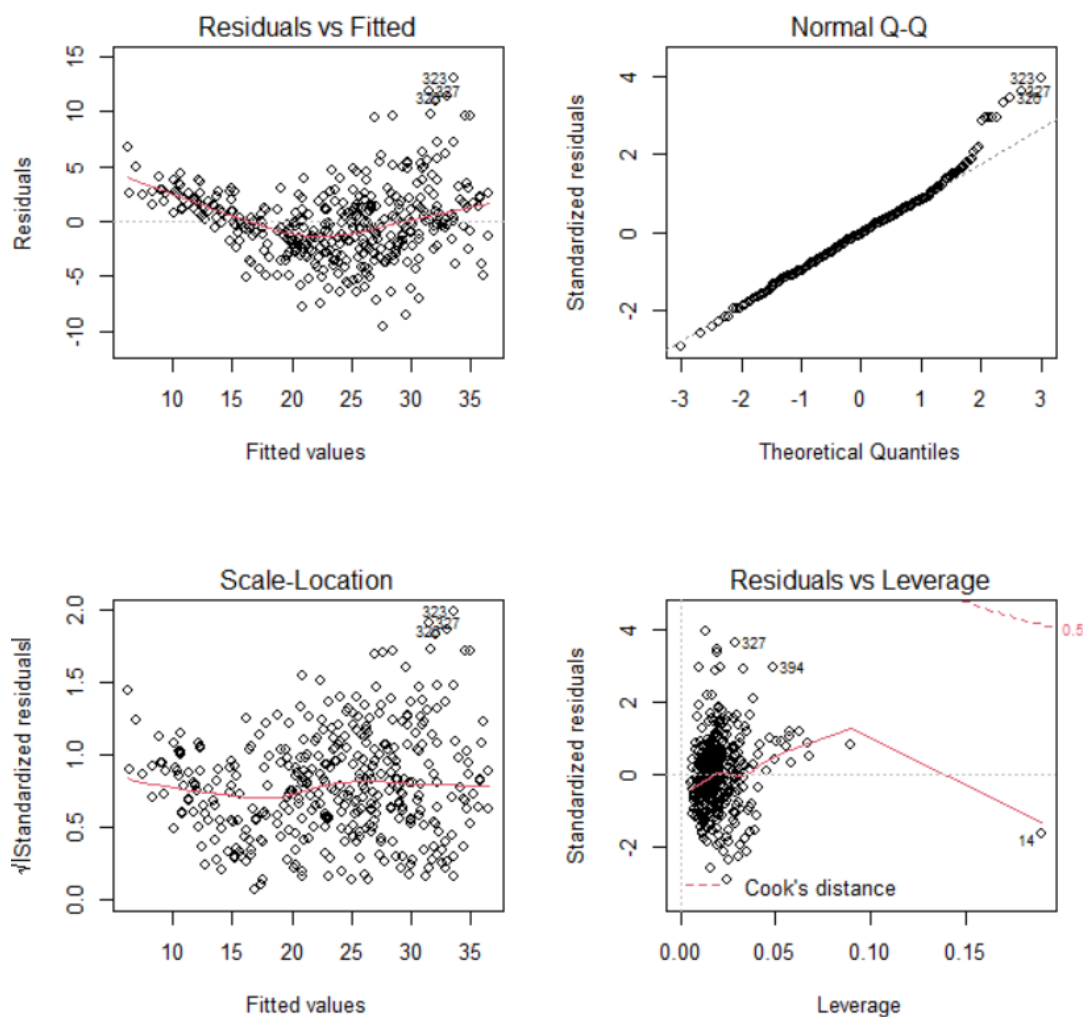
F-statistic є досить великою, тобто набагато більша за 1 з малим p-value, що свідчить проти нульової гіпотези про те що всі коефіцієнти регресії є нульовими, тобто є зв'язок між залежною змінною та предикторами.

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.

```

Call:
lm(formula = mpg ~ displacement * origin + acceleration * horsepower,
    data = autos)

Residuals:
    Min       1Q   Median       3Q      Max
-13.0588  -2.5884  -0.1985   2.0923  16.1532

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   24.753071    3.335079   7.422 7.46e-13 ***
displacement  -0.017668    0.010454  -1.690  0.09182  .
origin         2.293684    1.050508   2.183  0.02961  *
acceleration   0.896371    0.200563   4.469 1.03e-05 ***
horsepower     0.093591    0.029427   3.180  0.00159  **
displacement:origin -0.011570    0.009365  -1.235  0.21741
acceleration:horsepower -0.014348    0.001908  -7.519 3.91e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.05 on 385 degrees of freedom
Multiple R-squared:  0.7349,    Adjusted R-squared:  0.7307
F-statistic: 177.9 on 6 and 385 DF,  p-value: < 2.2e-16

```

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 , \sqrt{X} . Використав для цього функцію anova().

```

Analysis of Variance Table

Model 1: mpg ~ horsepower
Model 2: mpg ~ horsepower + I(log(horsepower))
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1     390 9385.9
2     389 7581.2  1     1804.7 92.601 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Analysis of Variance Table

Model 1: mpg ~ horsepower
Model 2: mpg ~ horsepower + I(horsepower^2)
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1     390 9385.9
2     389 7442.0  1     1943.9 101.61 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Analysis of Variance Table

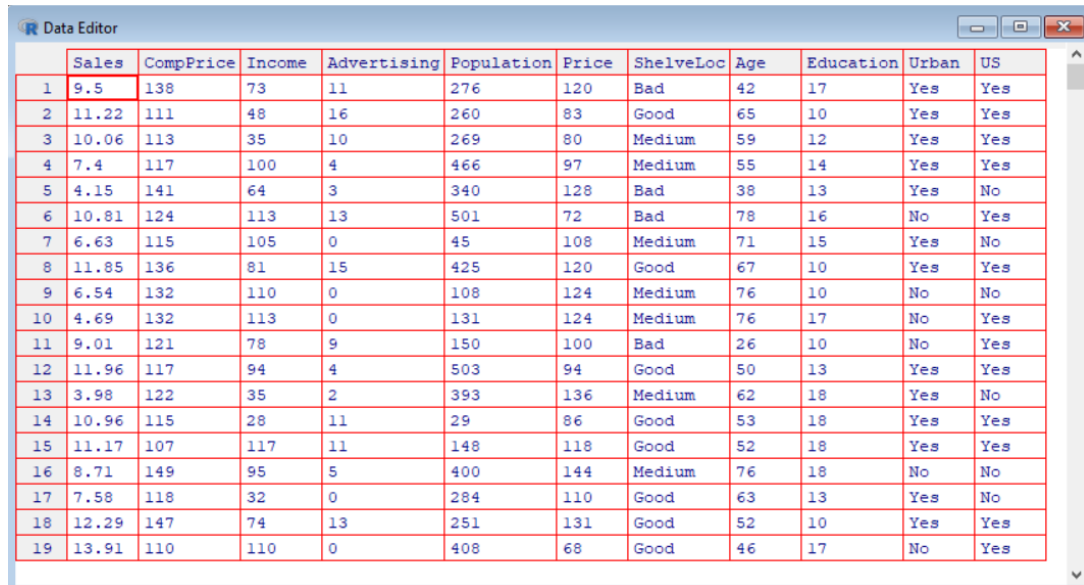
Model 1: mpg ~ horsepower
Model 2: mpg ~ horsepower + I(sqrt(horsepower))
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1     390 9385.9
2     389 7502.2  1     1883.7 97.672 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

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.



	Sales	CompPrice	Income	Advertising	Population	Price	ShelveLoc	Age	Education	Urban	US
1	9.5	138	73	11	276	120	Bad	42	17	Yes	Yes
2	11.22	111	48	16	260	83	Good	65	10	Yes	Yes
3	10.06	113	35	10	269	80	Medium	59	12	Yes	Yes
4	7.4	117	100	4	466	97	Medium	55	14	Yes	Yes
5	4.15	141	64	3	340	128	Bad	38	13	Yes	No
6	10.81	124	113	13	501	72	Bad	78	16	No	Yes
7	6.63	115	105	0	45	108	Medium	71	15	Yes	No
8	11.85	136	81	15	425	120	Good	67	10	Yes	Yes
9	6.54	132	110	0	108	124	Medium	76	10	No	No
10	4.69	132	113	0	131	124	Medium	76	17	No	Yes
11	9.01	121	78	9	150	100	Bad	26	10	No	Yes
12	11.96	117	94	4	503	94	Good	50	13	Yes	Yes
13	3.98	122	35	2	393	136	Medium	62	18	Yes	No
14	10.96	115	28	11	29	86	Good	53	18	Yes	Yes
15	11.17	107	117	11	148	118	Good	52	18	Yes	Yes
16	8.71	149	95	5	400	144	Medium	76	18	No	No
17	7.58	118	32	0	284	110	Good	63	13	Yes	No
18	12.29	147	74	13	251	131	Good	52	10	Yes	Yes
19	13.91	110	110	0	408	68	Good	46	17	No	Yes

A data frame with 400 observations on the following 11 variables.

Sales

Unit sales (in thousands) at each location

CompPrice

Price charged by competitor at each location

Income

Community income level (in thousands of dollars)

Advertising

Local advertising budget for company at each location (in thousands of dollars)

Population

Population size in region (in thousands)

Price

Price company charges for car seats at each site

ShelveLoc

A factor with levels Bad, Good and Medium indicating the quality of the shelving location for the car seats at each site

Age

Average age of the local population

Education

Education level at each location

Urban

A factor with levels No and Yes to indicate whether the store is in an urban or rural location

US

A factor with levels No and Yes to indicate whether the store is in the US or not

3.1

```
> lm.fit = lm(Sales~Price+Urban+US, data=Carseats)
> summary(lm.fit)

Call:
lm(formula = Sales ~ Price + Urban + US, data = Carseats)

Residuals:
    Min       1Q   Median       3Q      Max
-6.9206 -1.6220 -0.0564  1.5786  7.0581

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.043469   0.651012  20.036 < 2e-16 ***
Price       -0.054459   0.005242 -10.389 < 2e-16 ***
UrbanYes    -0.021916   0.271650  -0.081  0.936
USYes       1.200573   0.259042   4.635 4.86e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.472 on 396 degrees of freedom
Multiple R-squared:  0.2393,    Adjusted R-squared:  0.2335
F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

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 р-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

$$\text{Sales} = 13.04 + -0.05 * \text{Price} + -0.02 * \text{UrbanYes} + 1.20 * \text{USYes}$$

3.4

Нульову гіпотезу можна відхилити для гіпотези Price та USYes, на основі колонки $Pr(> |t|)$.

3.5 Нова модель

```
> lm.fit2 = lm(Sales ~ Price + US, data=Carseats)
> summary(lm.fit2)

Call:
lm(formula = Sales ~ Price + US, data = Carseats)

Residuals:
    Min       1Q   Median       3Q      Max
-6.9269 -1.6286 -0.0574  1.5766  7.0515

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.03079    0.63098   20.652 < 2e-16 ***
Price        -0.05448    0.00523  -10.416 < 2e-16 ***
USYes         1.19964    0.25846   4.641 4.71e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.469 on 397 degrees of freedom
Multiple R-squared:  0.2393,    Adjusted R-squared:  0.2354
F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

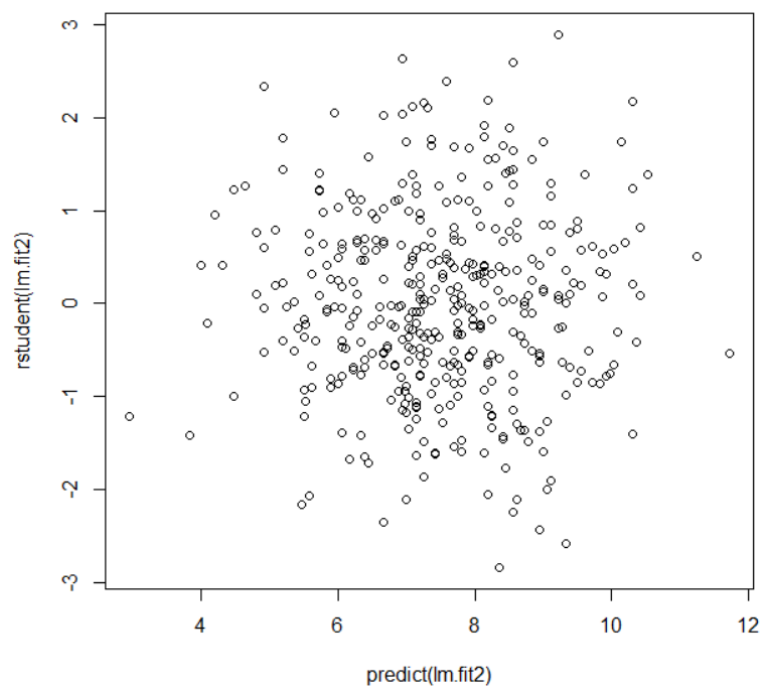
3.6

Зважаючи на значення RSE та R^2 можна стверджувати, що обидві моделі добре підходять для даних. Проте друга модель трошки краща.

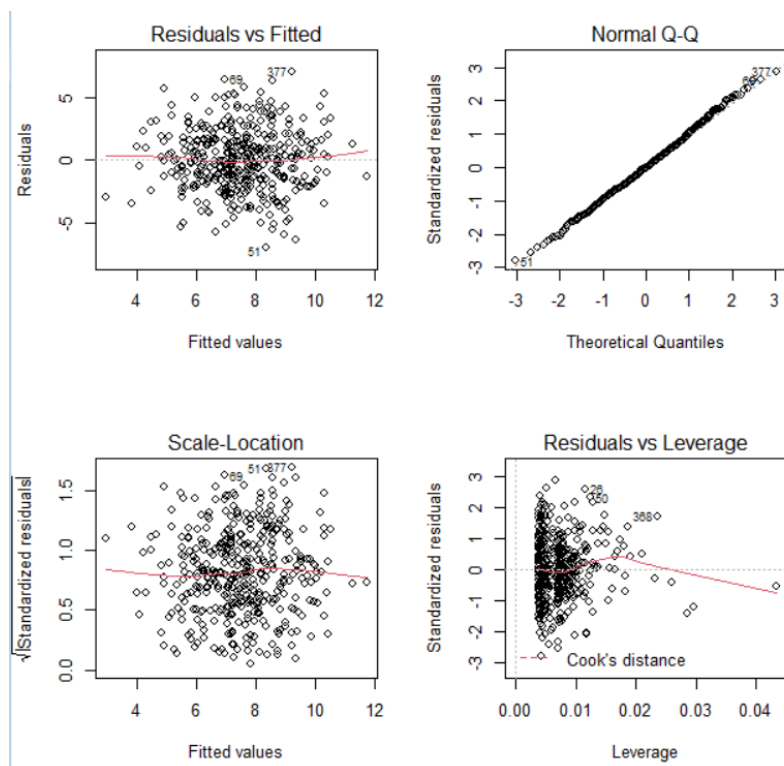
3.7

```
> confint(lm.fit2)
              2.5 %      97.5 %
(Intercept) 11.79032020 14.27126531
Price        -0.06475984 -0.04419543
USYes         0.69151957  1.70776632
```

3.8



Усі Студентифіковані залишки, обмежені від -3 до 3, тому з лінійної регресії не впливають 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 .

```
set.seed(1)
x = rnorm(100)
y = 2*x+rnorm(100)
```

4.1

Побудовано просту лінійну регресію y на x без β_0 .

```
Call:
lm(formula = y ~ x + 0)

Residuals:
    Min       1Q   Median       3Q      Max
-1.9154 -0.6472 -0.1771  0.5056  2.3109

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
x    1.9939     0.1065   18.73  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9586 on 99 degrees of freedom
Multiple R-squared:  0.7798,    Adjusted R-squared:  0.7776
F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

З наведеного вище p-value з t-статистики, яке є дуже малим (майже нульовим), можна зробити висновок про відкидання нульової гіпотези ($H_0: \beta = 0$).

4.2

Побудовано просту лінійну регресію x на y без β_0 .

```
Call:
lm(formula = x ~ y + 0)

Residuals:
    Min       1Q   Median       3Q      Max
-0.8699 -0.2368  0.1030  0.2858  0.8938

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
y  0.39111    0.02089    18.73  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4246 on 99 degrees of freedom
Multiple R-squared:  0.7798,    Adjusted R-squared:  0.7776
F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

З наведеного вище p-value з t-статистики, бачимо аналогічний результат з попередньою моделлю, тобто ми відкидаємо нульову гіпотезу ($H_0: \beta = 0$).

4.3

```
"Correlation x, y: 0.8822902418138"
```

Можемо бачити досить тісну кореляцію змінних y та x . Про це свідчить і той факт, $y=2x+\epsilon$ може бути розписане через $x=0.5*(y-\epsilon)$.

4.4

```
print(paste('T-statistic: ',
(sqrt(length(x)-1) * sum(x*y)) / (sqrt(sum(x*x) * sum(y*y) - (sum(x*y))^2))))
```

```
"T-statistic: 18.7259319374486"
```

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

$$\frac{(\sqrt{n-1}) \sum_{i=1}^n x_i y_i}{\sqrt{(\sum_{i=1}^n x_i^2)(\sum_{i'=1}^n y_{i'}^2) - (\sum_{i'=1}^n x_{i'} y_{i'})^2}}$$

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 $t \text{ value}=18.73$.

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 β_0 як для x на y , так й для y на x .

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.03769261	0.09698729	-0.3886346	6.983896e-01
x	1.99893961	0.10772703	18.5555993	7.723851e-34
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.03880394	0.04266144	0.9095787	3.652764e-01
y	0.38942451	0.02098690	18.5555993	7.723851e-34

Бачимо, що як і з моделями без коефіцієнта β_0 t -статистика для обох моделей є однаковою з $t \text{ value}=18.56$. При чому варто наголосити, що моделі з коефіцієнтом β_0 мають інше значення $t \text{ value}$ порівнюючи з попередніми моделями.

5. Знову розглянемо просту лінійну регресію без коефіцієнта β_0 .

5.1

$$\beta = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i'=1}^n x_{i'}^2}$$

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

$$\frac{\sum_{i=1}^n x_i y_i}{\sum_{i'=1}^n x_{i'}^2} = \frac{\sum_{i=1}^n y_i x_i}{\sum_{i'=1}^n y_{i'}^2} \Rightarrow \sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i^2$$

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

```
> set.seed(1)
> x=1:100
> y=2*x+rnorm(100)
> sum(x^2)
[1] 338350
> sum(y^2)
[1] 1355530
```

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

```
> val = lm(y~x+0)
> val$coefficients
      x
2.001514
> val = lm(x~y+0)
> val$coefficients
      y
0.4995922
```

З результатів видно що вони різні

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

```
> x = 1:100 + rnorm(100)
> y = x
> sum(x^2)
[1] 338194.8
> sum(y^2)
[1] 338194.8
```

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

```
> val = lm(y~x+0)
> val$coefficients
x
1
>
> val = lm(x~y+0)
> val$coefficients
y
1
```


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

6.1-6.3

Створено вектор x та eps з використанням функції `rnorm()`. З них побудовано y відповідно до моделі $y = -1 + 0,5X + \epsilon$.

```
# 6
set.seed(1)

# 6.1
x = rnorm(100)

print(x)
cat("\n")

# 6.2
eps = rnorm(100, 0, 0.25)
print(eps)

# 6.3
cat("\n")
y = -1 + 0.5*x + eps
```

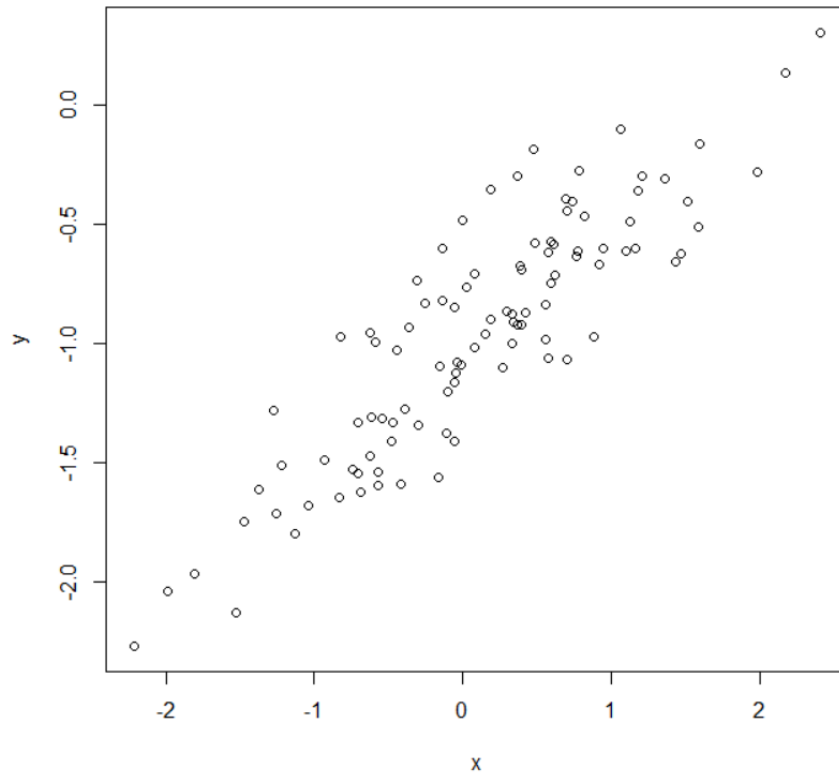
Як бачимо довжина вектора y - 100, $\beta_0 = -1$, а $\beta_1 = 0.5$.

```
"Vector y length: 100"
"Beta_0 = -1, Beta_1 = 0.5"
```

6.4

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

```
"Correlation between x and y: 0.8822902418138"
```



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 .

```
Call:
lm(formula = y ~ x)

Residuals:
    Min       1Q   Median       3Q      Max
-0.46921 -0.15344 -0.03487  0.13485  0.58654

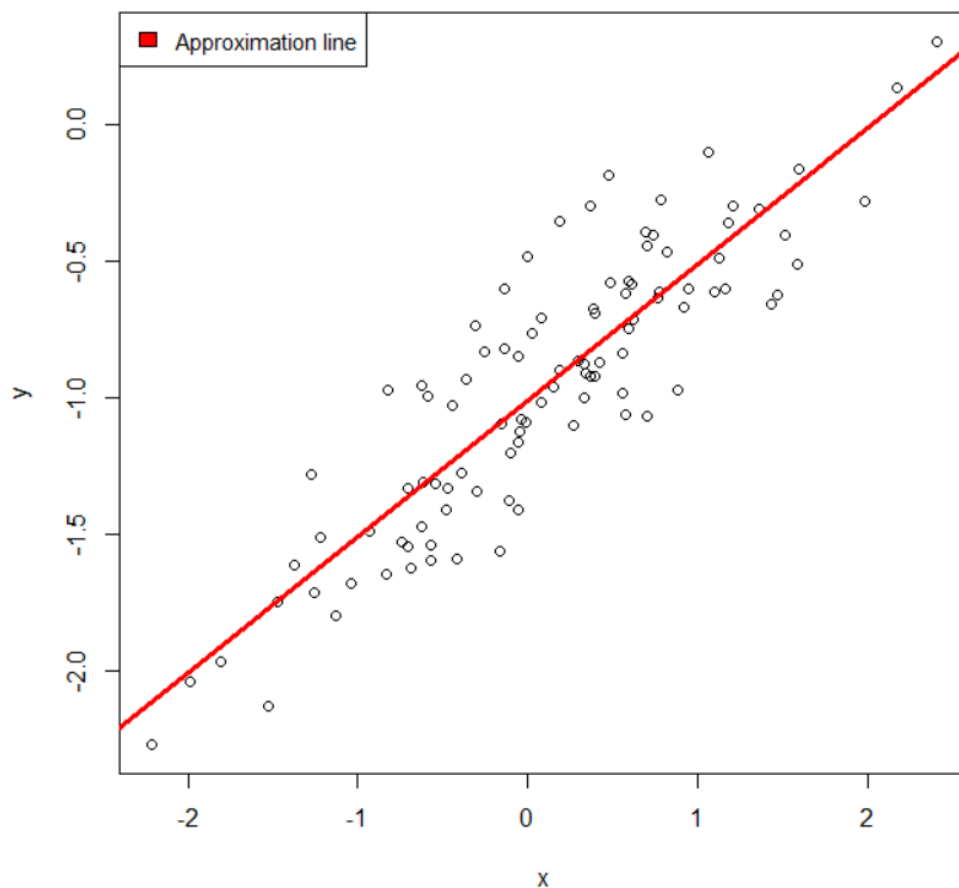
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.00942    0.02425  -41.63  <2e-16 ***
x             0.49973    0.02693   18.56  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2407 on 98 degrees of freedom
Multiple R-squared:  0.7784,    Adjusted R-squared:  0.7762
F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

Взявши до уваги аналіз нашої моделі впливає, 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 розсіювання.



6.7

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

```

Call:
lm(formula = y ~ x + I(x^2))

Residuals:
    Min       1Q   Median       3Q      Max
-0.4913 -0.1563 -0.0322  0.1451  0.5675

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.98582    0.02941  -33.516  <2e-16 ***
x             0.50429    0.02700   18.680  <2e-16 ***
I(x^2)       -0.02973    0.02119   -1.403    0.164
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2395 on 97 degrees of freedom
Multiple R-squared:  0.7828,    Adjusted R-squared:  0.7784
F-statistic: 174.8 on 2 and 97 DF,  p-value: < 2.2e-16

Analysis of Variance Table

Model 1: y ~ x
Model 2: y ~ x + I(x^2)
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     98 5.6772
2     97 5.5643  1   0.11291 1.9682 0.1638

```

Бачимо, що значення β_0 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 ϵ до 0.05).

```

Call:
lm(formula = y2 ~ x)

Residuals:
    Min       1Q   Median       3Q      Max
-0.145706 -0.024115 -0.002266  0.032462  0.132079

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.998632   0.005235  -190.75   <2e-16 ***
x             0.501058   0.005815   86.17   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05197 on 98 degrees of freedom
Multiple R-squared:  0.987,    Adjusted R-squared:  0.9868
F-statistic: 7425 on 1 and 98 DF,  p-value: < 2.2e-16

```

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 ϵ до 0.5).

```

Call:
lm(formula = y3 ~ x)

Residuals:
    Min       1Q   Median       3Q      Max
-1.25813 -0.27262 -0.01888  0.33644  0.93944

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.97117   0.05014  -19.369   < 2e-16 ***
x             0.47216   0.05569   8.478  2.4e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4977 on 98 degrees of freedom
Multiple R-squared:  0.4231,    Adjusted R-squared:  0.4172
F-statistic: 71.87 on 1 and 98 DF,  p-value: 2.4e-13

```

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.

	2.5 %	97.5 %
(Intercept)	-1.0575402	-0.9613061
x	0.4462897	0.5531801
	2.5 %	97.5 %
(Intercept)	-1.0090206	-0.9882425
x	0.4895188	0.5125978
	2.5 %	97.5 %
(Intercept)	-1.070670	-0.8716647
x	0.361636	0.5826779

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

7. Зосередимося на проблемі колінеарності.

7.1

Форма лінійної моделі та коефіцієнти регресії

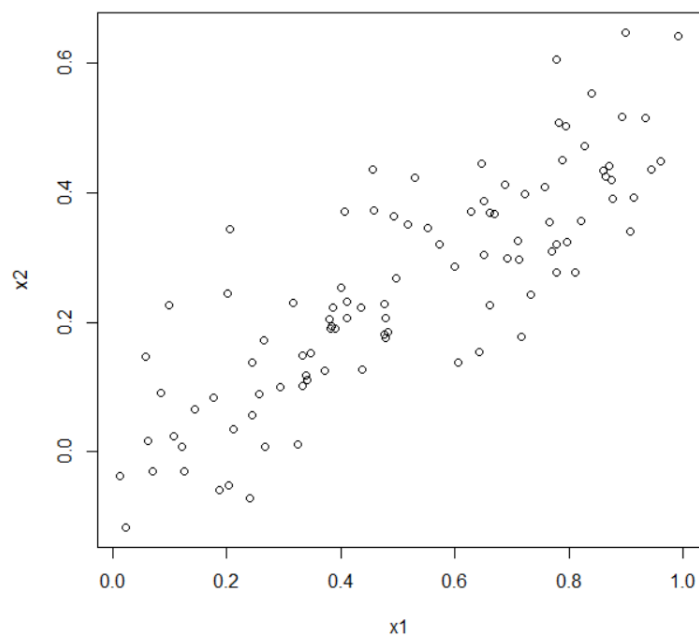
$$Y = 2 + 2X_1 + 0.3X_2 + \varepsilon$$

$$\beta_0 = 2, \beta_1 = 2, \beta_3 = 0.3$$

7.2

```
> cor(x1, x2)
[1] 0.8351212
```

Кореляція між x_1 та x_2 та діаграма розсіювання



7.3

```
> lm.fit = lm(y~x1+x2)
> summary(lm.fit)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min       1Q   Median       3Q      Max
-2.8311 -0.7273 -0.0537  0.6338  2.3359

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
x1             1.4396     0.7212   1.996  0.0487 *
x2             1.0097     1.1337   0.891  0.3754
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.056 on 97 degrees of freedom
Multiple R-squared:  0.2088,    Adjusted R-squared:  0.1925
F-statistic: 12.8 on 2 and 97 DF,  p-value: 1.164e-05
```

$$\beta_0 = 2.1305, \beta_1 = 1.4396, \beta_2 = 1.0097$$

Можемо побачити що коефіцієнт β_0 близький до реального значення і значення p надзвичайно мале що вказує на те, що модель підібрала його коректною.

Значення β_1 достатньо близьке до реального його значення ρ може бути достатньо для відхилення нульової гіпотези.

Значення β_3 найгірше, а найвище значення ρ означає, що ми можемо прийняти 0 гіпотезу.

7.4

```
> lm.fit = lm(y~x1)
> summary(lm.fit)

Call:
lm(formula = y ~ x1)

Residuals:
    Min       1Q   Median       3Q      Max
-2.89495 -0.66874 -0.07785  0.59221  2.45560

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.1124      0.2307   9.155 8.27e-15 ***
x1             1.9759      0.3963   4.986 2.66e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.055 on 98 degrees of freedom
Multiple R-squared:  0.2024,    Adjusted R-squared:  0.1942
F-statistic: 24.86 on 1 and 98 DF,  p-value: 2.661e-06
```

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

7.5

```
> lm.fit = lm(y~x2)
> summary(lm.fit)

Call:
lm(formula = y ~ x2)

Residuals:
    Min       1Q   Median       3Q      Max
-2.62687 -0.75156 -0.03598  0.72383  2.44890

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.3899      0.1949  12.26 < 2e-16 ***
x2             2.8996      0.6330   4.58 1.37e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.072 on 98 degrees of freedom
Multiple R-squared:  0.1763,    Adjusted R-squared:  0.1679
F-statistic: 20.98 on 1 and 98 DF,  p-value: 1.366e-05
```


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

7.6

Ні, оскільки x_1 та x_2 мають колінеарність, важко відрізнити їх вплив, коли вони регресуються разом. Коли вони регресуються окремо, лінійна залежність між y і кожним предиктором визначаються більш чітко.

7.7

```
> x1 = c(x1,0.1)
> x2 = c(x2,0.8)
> y = c(y,6)
> lm.fit1 = lm(y~x1+x2)
> summary(lm.fit1)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min       1Q   Median       3Q      Max
-2.73348 -0.69318 -0.05263  0.66385  2.30619

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2267     0.2314   9.624 7.91e-16 ***
x1             0.5394     0.5922   0.911  0.36458
x2             2.5146     0.8977   2.801  0.00614 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.075 on 98 degrees of freedom
Multiple R-squared:  0.2188,    Adjusted R-squared:  0.2029
F-statistic: 13.72 on 2 and 98 DF,  p-value: 5.564e-06
```

```

> lm.fit2 = lm(y~x1)
> summary(lm.fit2)

Call:
lm(formula = y ~ x1)

Residuals:
    Min       1Q   Median       3Q      Max
-2.8897 -0.6556 -0.0909  0.5682  3.5665

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.2569     0.2390   9.445 1.78e-15 ***
x1           1.7657     0.4124   4.282 4.29e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.111 on 99 degrees of freedom
Multiple R-squared:  0.1562,    Adjusted R-squared:  0.1477
F-statistic: 18.33 on 1 and 99 DF,  p-value: 4.295e-05


> lm.fit3 = lm(y~x2)
> summary(lm.fit3)

Call:
lm(formula = y ~ x2)

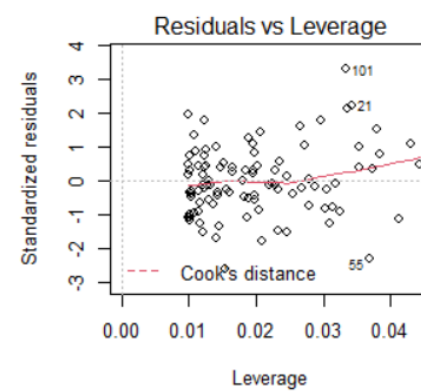
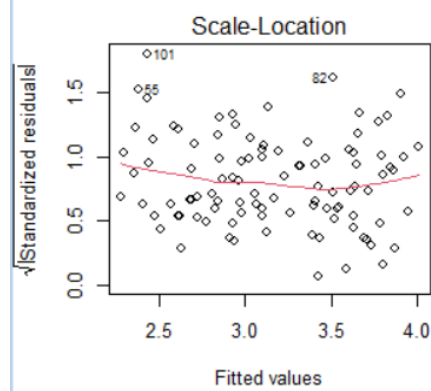
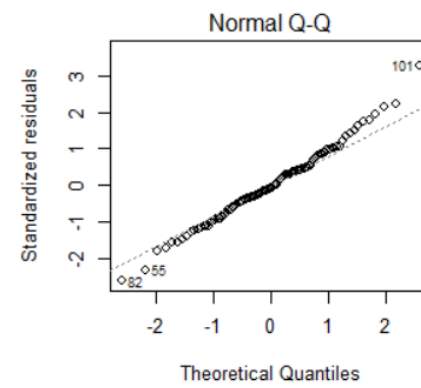
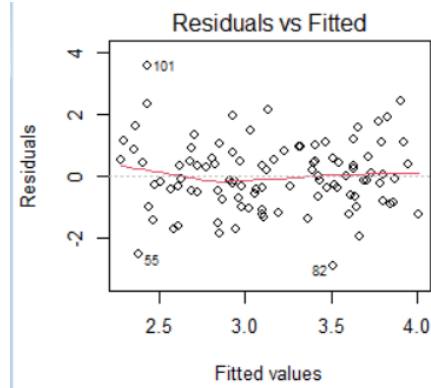
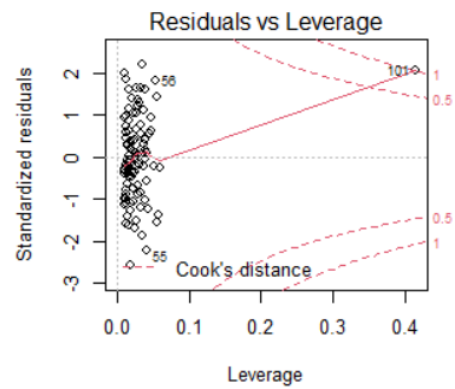
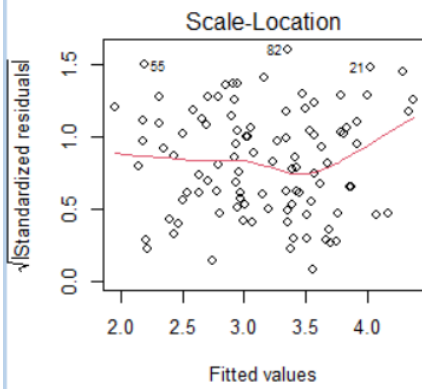
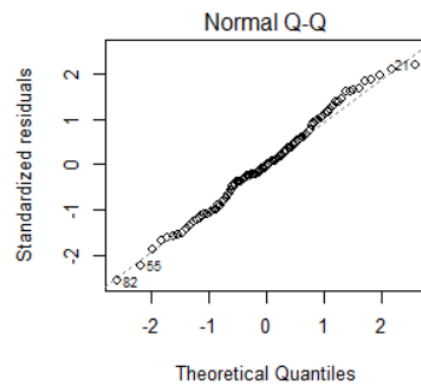
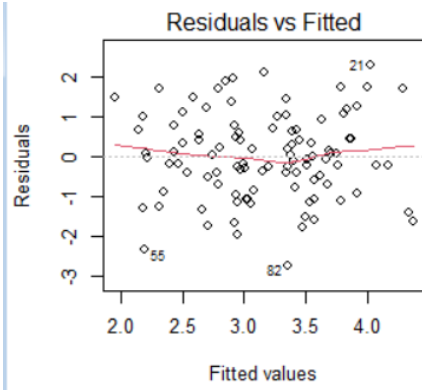
Residuals:
    Min       1Q   Median       3Q      Max
-2.64729 -0.71021 -0.06899  0.72699  2.38074

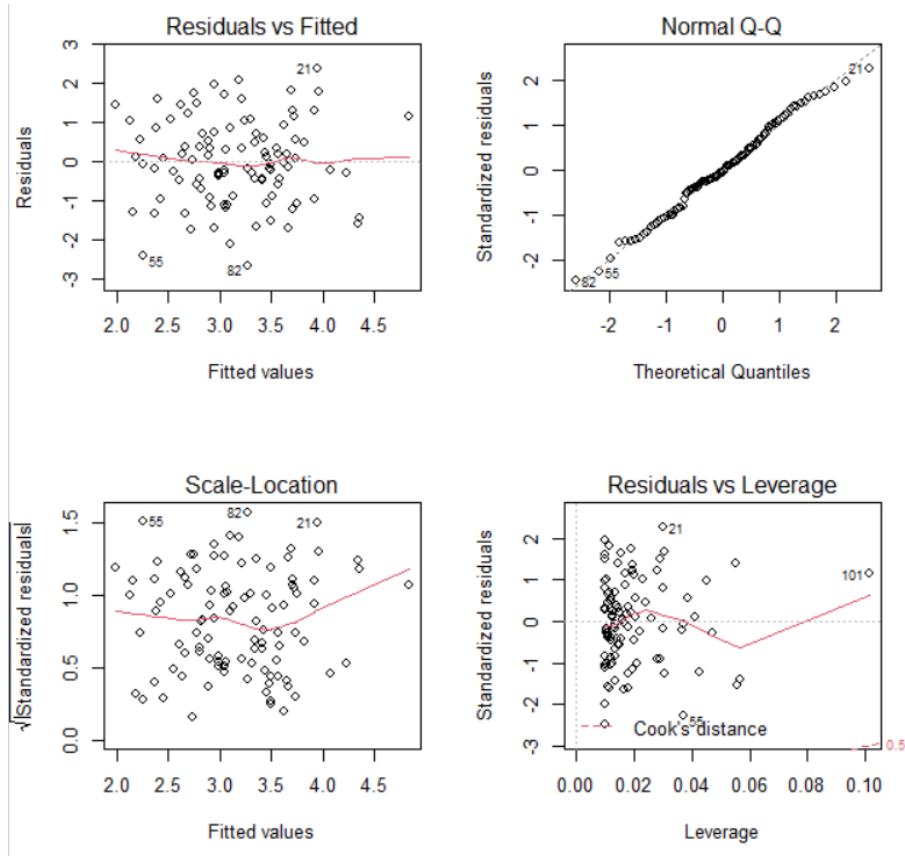
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.3451     0.1912  12.264 < 2e-16 ***
x2           3.1190     0.6040   5.164 1.25e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.074 on 99 degrees of freedom
Multiple R-squared:  0.2122,    Adjusted R-squared:  0.2042
F-statistic: 26.66 on 1 and 99 DF,  p-value: 1.253e-06

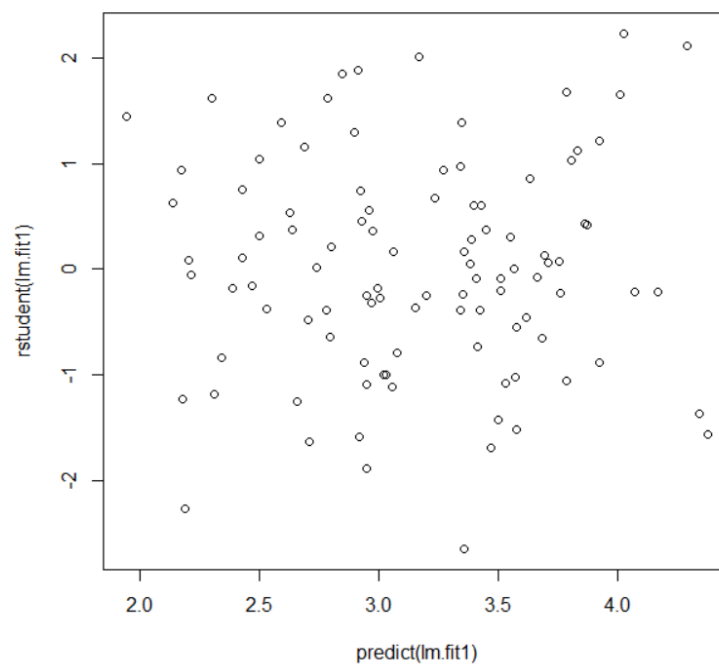
```

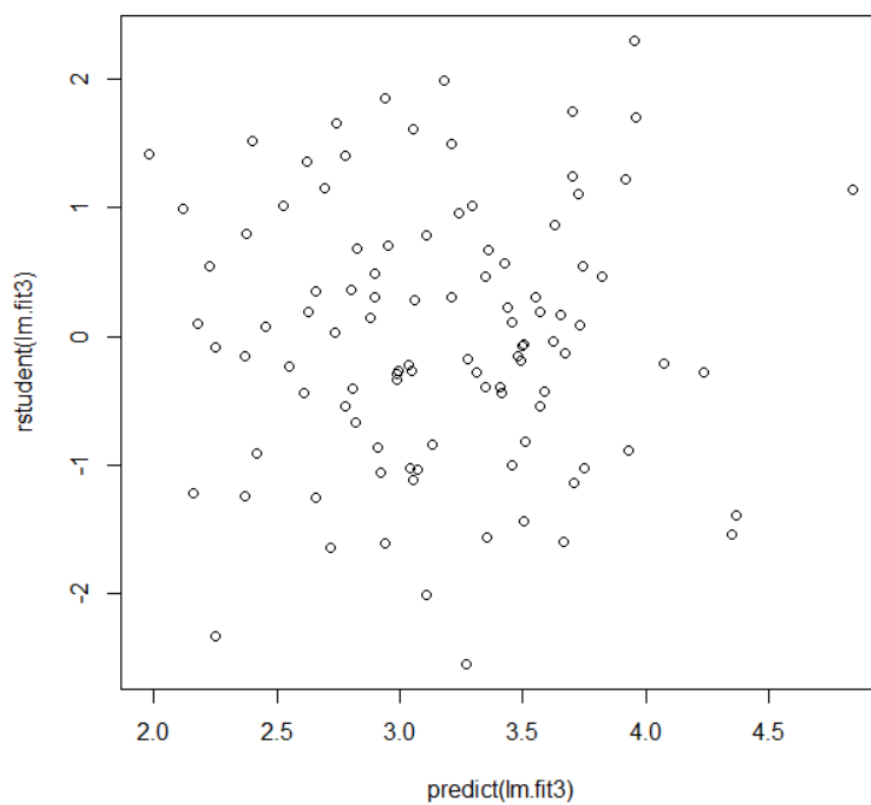
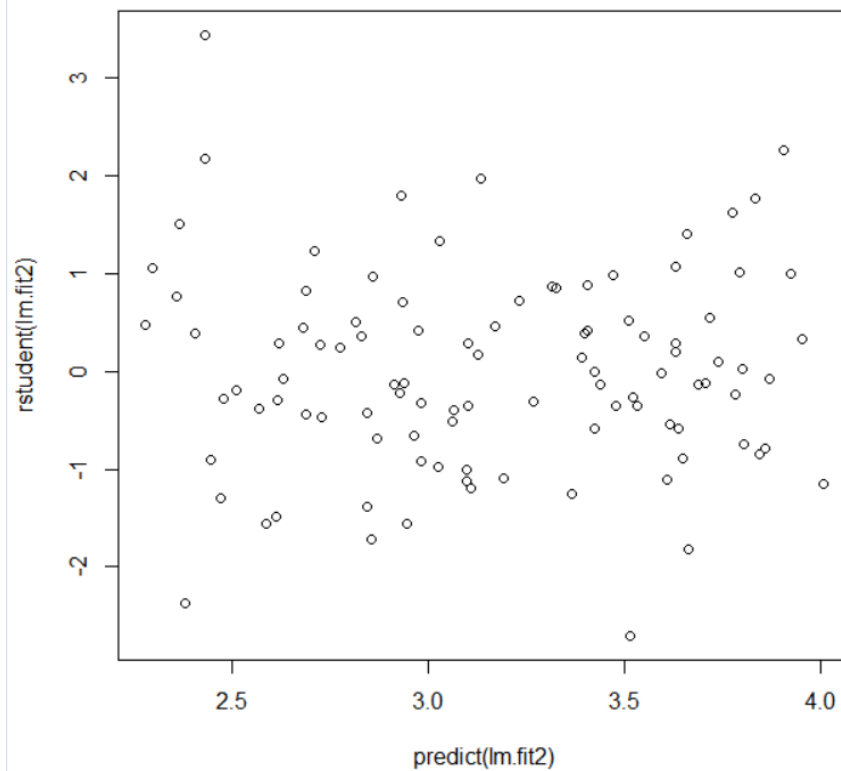
У першій моделі вона зсуває x_1 до статистично незначущої та x_2 до статистично значущої від зміни р-значень між двома лінійними регресіями.





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 $|3|$, за винятком другої лінійної регресії: $y \sim x_1$.

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

crim	zn	indus	chas
Min. : 0.00632	Min. : 0.00	Min. : 0.46	Min. : 0.00000
1st Qu.: 0.08205	1st Qu.: 0.00	1st Qu.: 5.19	1st Qu.: 0.00000
Median : 0.25651	Median : 0.00	Median : 9.69	Median : 0.00000
Mean : 3.61352	Mean : 11.36	Mean : 11.14	Mean : 0.06917
3rd Qu.: 3.67708	3rd Qu.: 12.50	3rd Qu.: 18.10	3rd Qu.: 0.00000
Max. : 88.97620	Max. : 100.00	Max. : 27.74	Max. : 1.00000
nox	rm	age	dis
Min. : 0.3850	Min. : 3.561	Min. : 2.90	Min. : 1.130
1st Qu.: 0.4490	1st Qu.: 5.886	1st Qu.: 45.02	1st Qu.: 2.100
Median : 0.5380	Median : 6.208	Median : 77.50	Median : 3.207
Mean : 0.5547	Mean : 6.285	Mean : 68.57	Mean : 3.795
3rd Qu.: 0.6240	3rd Qu.: 6.623	3rd Qu.: 94.08	3rd Qu.: 5.188
Max. : 0.8710	Max. : 8.780	Max. : 100.00	Max. : 12.127
rad	tax	ptratio	black
Min. : 1.000	Min. : 187.0	Min. : 12.60	Min. : 0.32
1st Qu.: 4.000	1st Qu.: 279.0	1st Qu.: 17.40	1st Qu.: 375.38
Median : 5.000	Median : 330.0	Median : 19.05	Median : 391.44
Mean : 9.549	Mean : 408.2	Mean : 18.46	Mean : 356.67
3rd Qu.: 24.000	3rd Qu.: 666.0	3rd Qu.: 20.20	3rd Qu.: 396.23
Max. : 24.000	Max. : 711.0	Max. : 22.00	Max. : 396.90
lstat	medv		
Min. : 1.73	Min. : 5.00		
1st Qu.: 6.95	1st Qu.: 17.02		
Median : 11.36	Median : 21.20		
Mean : 12.65	Mean : 22.53		
3rd Qu.: 16.95	3rd Qu.: 25.00		
Max. : 37.97	Max. : 50.00		

Загальна характеристика даних Boston

8.1

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

```

Call:
lm(formula = crim ~ zn)

Residuals:
    Min       1Q   Median       3Q      Max
-4.429 -4.222 -2.620  1.250  84.523

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.45369    0.41722   10.675 < 2e-16 ***
zn          -0.07393    0.01609   -4.594 5.51e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.435 on 504 degrees of freedom
Multiple R-squared:  0.04019,    Adjusted R-squared:  0.03828
F-statistic: 21.1 on 1 and 504 DF,  p-value: 5.506e-06


Call:
lm(formula = crim ~ indus)

Residuals:
    Min       1Q   Median       3Q      Max
-11.972  -2.698  -0.736    0.712   81.813

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.06374    0.66723   -3.093  0.00209 **
indus        0.50978    0.05102    9.991 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.866 on 504 degrees of freedom
Multiple R-squared:  0.1653,    Adjusted R-squared:  0.1637
F-statistic: 99.82 on 1 and 504 DF,  p-value: < 2.2e-16

```

Модель лінійної регресії для показника zn та indus

```

Call:
lm(formula = crim ~ chas)

Residuals:
    Min       1Q   Median       3Q      Max
-3.738 -3.661 -3.435  0.018 85.232

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.7444    0.3961   9.453  <2e-16 ***
chas         -1.8928    1.5061  -1.257   0.209
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.597 on 504 degrees of freedom
Multiple R-squared:  0.003124, Adjusted R-squared:  0.001146
F-statistic: 1.579 on 1 and 504 DF, p-value: 0.2094

Call:
lm(formula = crim ~ nox)

Residuals:
    Min       1Q   Median       3Q      Max
-12.371 -2.738 -0.974  0.559 81.728

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -13.720    1.699  -8.073 5.08e-15 ***
nox           31.249    2.999  10.419 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.81 on 504 degrees of freedom
Multiple R-squared:  0.1772, Adjusted R-squared:  0.1756
F-statistic: 108.6 on 1 and 504 DF, p-value: < 2.2e-16

```

Модель лінійної регресії для показника chas та nox


```

Call:
lm(formula = crim ~ rm)

Residuals:
    Min       1Q   Median       3Q      Max
-6.604 -3.952 -2.654  0.989 87.197

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.482      3.365   6.088 2.27e-09 ***
rm          -2.684      0.532  -5.045 6.35e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.401 on 504 degrees of freedom
Multiple R-squared:  0.04807, Adjusted R-squared:  0.04618
F-statistic: 25.45 on 1 and 504 DF, p-value: 6.347e-07


Call:
lm(formula = crim ~ age)

Residuals:
    Min       1Q   Median       3Q      Max
-6.789 -4.257 -1.230  1.527 82.849

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.77791      0.94398  -4.002 7.22e-05 ***
age          0.10779      0.01274   8.463 2.85e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.057 on 504 degrees of freedom
Multiple R-squared:  0.1244, Adjusted R-squared:  0.1227
F-statistic: 71.62 on 1 and 504 DF, p-value: 2.855e-16

```

Модель лінійної регресії для показника `rm` та `age`

```

Call:
lm(formula = crim ~ dis)

Residuals:
    Min       1Q   Median       3Q      Max
-6.708 -4.134 -1.527  1.516 81.674

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.4993     0.7304  13.006  <2e-16 ***
dis          -1.5509     0.1683  -9.213  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.965 on 504 degrees of freedom
Multiple R-squared:  0.1441,    Adjusted R-squared:  0.1425
F-statistic: 84.89 on 1 and 504 DF,  p-value: < 2.2e-16

Call:
lm(formula = crim ~ rad)

Residuals:
    Min       1Q   Median       3Q      Max
-10.164 -1.381 -0.141  0.660 76.433

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.28716     0.44348  -5.157 3.61e-07 ***
rad           0.61791     0.03433  17.998  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.718 on 504 degrees of freedom
Multiple R-squared:  0.3913,    Adjusted R-squared:  0.39
F-statistic: 323.9 on 1 and 504 DF,  p-value: < 2.2e-16

```

Модель лінійної регресії для показника dis та rad

```

Call:
lm(formula = crim ~ tax)

Residuals:
    Min       1Q   Median       3Q      Max
-12.513  -2.738  -0.194   1.065  77.696

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.528369   0.815809  -10.45  <2e-16 ***
tax           0.029742   0.001847   16.10  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.997 on 504 degrees of freedom
Multiple R-squared:  0.3396,    Adjusted R-squared:  0.3383
F-statistic: 259.2 on 1 and 504 DF,  p-value: < 2.2e-16

Call:
lm(formula = crim ~ ptratio)

Residuals:
    Min       1Q   Median       3Q      Max
-7.654  -3.985  -1.912   1.825  83.353

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.6469    3.1473  -5.607 3.40e-08 ***
ptratio       1.1520     0.1694   6.801 2.94e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.24 on 504 degrees of freedom
Multiple R-squared:  0.08407,    Adjusted R-squared:  0.08225
F-statistic: 46.26 on 1 and 504 DF,  p-value: 2.943e-11

```

Модель лінійної регресії для показника tax та ptratio

```

Call:
lm(formula = crim ~ black)

Residuals:
    Min       1Q   Median       3Q      Max
-13.756  -2.299  -2.095  -1.296   86.822

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.553529   1.425903  11.609  <2e-16 ***
black       -0.036280   0.003873   -9.367  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.946 on 504 degrees of freedom
Multiple R-squared:  0.1483,    Adjusted R-squared:  0.1466
F-statistic: 87.74 on 1 and 504 DF,  p-value: < 2.2e-16

Call:
lm(formula = crim ~ lstat)

Residuals:
    Min       1Q   Median       3Q      Max
-13.925  -2.822  -0.664   1.079   82.862

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.33054    0.69376   -4.801 2.09e-06 ***
lstat        0.54880    0.04776  11.491 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.664 on 504 degrees of freedom
Multiple R-squared:  0.2076,    Adjusted R-squared:  0.206
F-statistic: 132 on 1 and 504 DF,  p-value: < 2.2e-16

```

Модель лінійної регресії для для показника black та lstat

```

Call:
lm(formula = crim ~ medv)

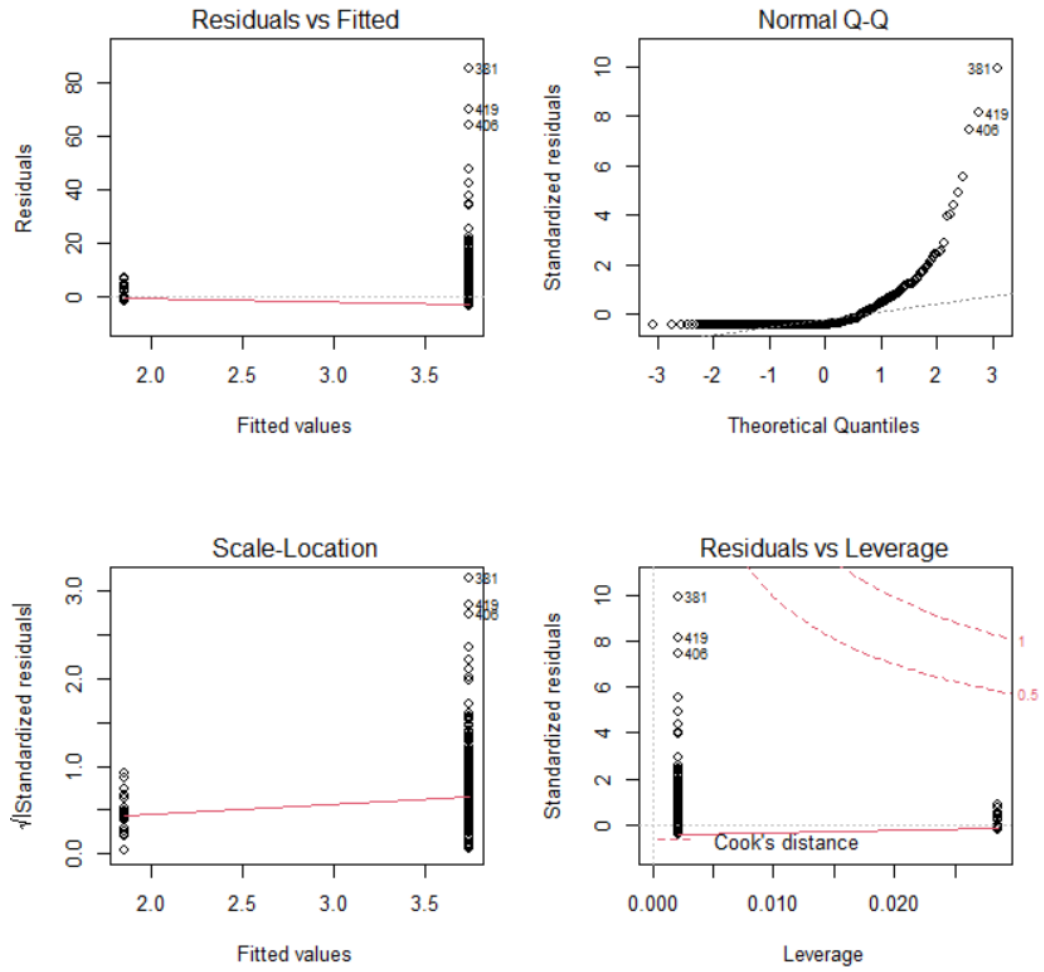
Residuals:
    Min       1Q   Median       3Q      Max
-9.071 -4.022 -2.343  1.298  80.957

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.79654    0.93419   12.63  <2e-16 ***
medv       -0.36316    0.03839   -9.46  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

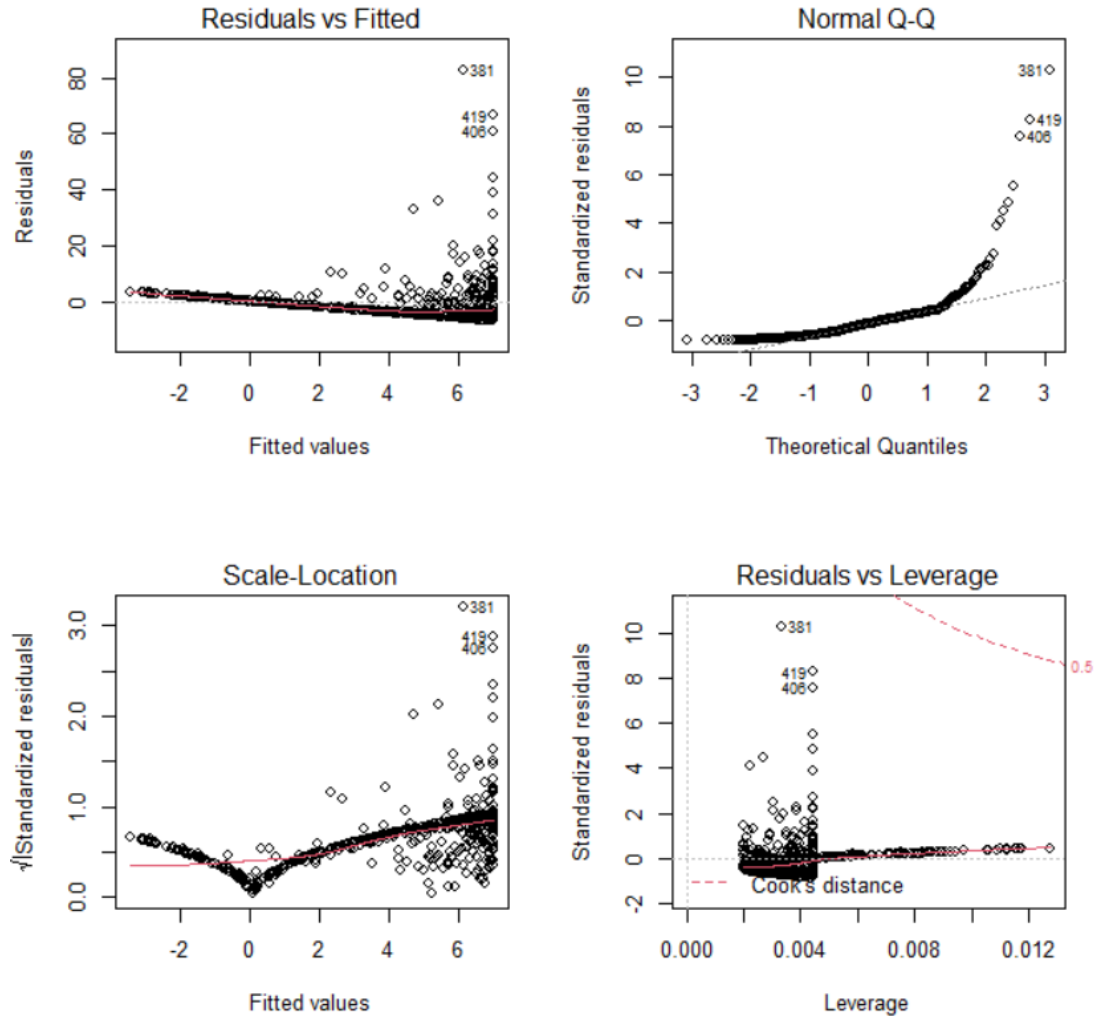
Residual standard error: 7.934 on 504 degrees of freedom
Multiple R-squared:  0.1508,    Adjusted R-squared:  0.1491
F-statistic: 89.49 on 1 and 504 DF,  p-value: < 2.2e-16

```

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.

```

Call:
lm(formula = crim ~ ., data = Boston)

Residuals:
    Min       1Q   Median       3Q      Max
-9.924 -2.120 -0.353  1.019  75.051

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.033228   7.234903   2.354 0.018949 *
zn           0.044855   0.018734   2.394 0.017025 *
indus       -0.063855   0.083407  -0.766 0.444294
chas        -0.749134   1.180147  -0.635 0.525867
nox        -10.313535   5.275536  -1.955 0.051152 .
rm           0.430131   0.612830   0.702 0.483089
age          0.001452   0.017925   0.081 0.935488
dis         -0.987176   0.281817  -3.503 0.000502 ***
rad          0.588209   0.088049   6.680 6.46e-11 ***
tax         -0.003780   0.005156  -0.733 0.463793
ptratio     -0.271081   0.186450  -1.454 0.146611
black       -0.007538   0.003673  -2.052 0.040702 *
lstat        0.126211   0.075725   1.667 0.096208 .
medv       -0.198887   0.060516  -3.287 0.001087 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.439 on 492 degrees of freedom
Multiple R-squared:  0.454,    Adjusted R-squared:  0.4396
F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16

```

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.

```

all_coeffs = c(coefficients(lm.zn)[2],
               coefficients(lm.indus)[2],
               coefficients(lm.chas)[2],
               coefficients(lm.nox)[2],
               coefficients(lm.rm)[2],
               coefficients(lm.age)[2],
               coefficients(lm.dis)[2],
               coefficients(lm.rad)[2],
               coefficients(lm.tax)[2],
               coefficients(lm.ptratio)[2],
               coefficients(lm.black)[2],
               coefficients(lm.lstat)[2],
               coefficients(lm.medv)[2])

print(all_coeffs)
cat("\n")
print(coefficients(lm.all)[2:14])

```

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

zn	indus	chas	nox	rm	age
-0.07393498	0.50977633	-1.89277655	31.24853120	-2.68405122	0.10778623
dis	rad	tax	ptratio	black	lstat
-1.55090168	0.61791093	0.02974225	1.15198279	-0.03627964	0.54880478
medv					
-0.36315992					
zn	indus	chas	nox	rm	
0.044855215	-0.063854824	-0.749133611	-10.313534912	0.430130506	
age	dis	rad	tax	ptratio	
0.001451643	-0.987175726	0.588208591	-0.003780016	-0.271080558	
black	lstat	medv			
-0.007537505	0.126211376	-0.198886821			

8.4

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

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


```

Call:
lm(formula = crim ~ poly(zn, 3))

Residuals:
    Min       1Q   Median       3Q      Max
-4.821 -4.614 -1.294   0.473  84.130

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.6135     0.3722   9.709 < 2e-16 ***
poly(zn, 3)1  -38.7498     8.3722  -4.628  4.7e-06 ***
poly(zn, 3)2   23.9398     8.3722   2.859  0.00442 **
poly(zn, 3)3  -10.0719     8.3722  -1.203  0.22954
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.372 on 502 degrees of freedom
Multiple R-squared:  0.05824,    Adjusted R-squared:  0.05261
F-statistic: 10.35 on 3 and 502 DF,  p-value: 1.281e-06


Call:
lm(formula = crim ~ poly(indus, 3))

Residuals:
    Min       1Q   Median       3Q      Max
-8.278 -2.514   0.054   0.764  79.713

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.614      0.330  10.950 < 2e-16 ***
poly(indus, 3)1  78.591     7.423  10.587 < 2e-16 ***
poly(indus, 3)2 -24.395     7.423  -3.286  0.00109 **
poly(indus, 3)3 -54.130     7.423  -7.292  1.2e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.423 on 502 degrees of freedom
Multiple R-squared:  0.2597,    Adjusted R-squared:  0.2552
F-statistic: 58.69 on 3 and 502 DF,  p-value: < 2.2e-16

```

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

```

Call:
lm(formula = crim ~ poly(nox, 3))

Residuals:
    Min       1Q   Median       3Q      Max
-9.110 -2.068 -0.255  0.739 78.302

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.6135     0.3216  11.237 < 2e-16 ***
poly(nox, 3)1  81.3720     7.2336  11.249 < 2e-16 ***
poly(nox, 3)2 -28.8286     7.2336  -3.985 7.74e-05 ***
poly(nox, 3)3 -60.3619     7.2336  -8.345 6.96e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.234 on 502 degrees of freedom
Multiple R-squared:  0.297,    Adjusted R-squared:  0.2928
F-statistic: 70.69 on 3 and 502 DF,  p-value: < 2.2e-16

Call:
lm(formula = crim ~ poly(rm, 3))

Residuals:
    Min       1Q   Median       3Q      Max
-18.485  -3.468  -2.221   -0.015   87.219

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.6135     0.3703   9.758 < 2e-16 ***
poly(rm, 3)1 -42.3794     8.3297  -5.088 5.13e-07 ***
poly(rm, 3)2  26.5768     8.3297   3.191 0.00151 **
poly(rm, 3)3  -5.5103     8.3297  -0.662 0.50858
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.33 on 502 degrees of freedom
Multiple R-squared:  0.06779,    Adjusted R-squared:  0.06222
F-statistic: 12.17 on 3 and 502 DF,  p-value: 1.067e-07

```

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

```

Call:
lm(formula = crim ~ poly(age, 3))

Residuals:
    Min       1Q   Median       3Q      Max
-9.762 -2.673 -0.516  0.019 82.842

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.6135     0.3485  10.368 < 2e-16 ***
poly(age, 3)1   68.1820     7.8397   8.697 < 2e-16 ***
poly(age, 3)2   37.4845     7.8397   4.781 2.29e-06 ***
poly(age, 3)3   21.3532     7.8397   2.724 0.00668 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.84 on 502 degrees of freedom
Multiple R-squared:  0.1742,    Adjusted R-squared:  0.1693
F-statistic: 35.31 on 3 and 502 DF,  p-value: < 2.2e-16

Call:
lm(formula = crim ~ poly(dis, 3))

Residuals:
    Min       1Q   Median       3Q      Max
-10.757 -2.588  0.031  1.267 76.378

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.6135     0.3259  11.087 < 2e-16 ***
poly(dis, 3)1 -73.3886     7.3315 -10.010 < 2e-16 ***
poly(dis, 3)2  56.3730     7.3315   7.689 7.87e-14 ***
poly(dis, 3)3 -42.6219     7.3315  -5.814 1.09e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.331 on 502 degrees of freedom
Multiple R-squared:  0.2778,    Adjusted R-squared:  0.2735
F-statistic: 64.37 on 3 and 502 DF,  p-value: < 2.2e-16

```

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

```

Call:
lm(formula = crim ~ poly(rad, 3))

Residuals:
    Min       1Q   Median       3Q      Max
-10.381  -0.412  -0.269   0.179   76.217

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.6135     0.2971  12.164 < 2e-16 ***
poly(rad, 3)1 120.9074     6.6824  18.093 < 2e-16 ***
poly(rad, 3)2  17.4923     6.6824   2.618 0.00912 **
poly(rad, 3)3   4.6985     6.6824   0.703 0.48231
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.682 on 502 degrees of freedom
Multiple R-squared:  0.4,    Adjusted R-squared:  0.3965
F-statistic: 111.6 on 3 and 502 DF,  p-value: < 2.2e-16

Call:
lm(formula = crim ~ poly(tax, 3))

Residuals:
    Min       1Q   Median       3Q      Max
-13.273  -1.389   0.046   0.536   76.950

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.6135     0.3047  11.860 < 2e-16 ***
poly(tax, 3)1 112.6458     6.8537  16.436 < 2e-16 ***
poly(tax, 3)2  32.0873     6.8537   4.682 3.67e-06 ***
poly(tax, 3)3  -7.9968     6.8537  -1.167  0.244
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.854 on 502 degrees of freedom
Multiple R-squared:  0.3689,    Adjusted R-squared:  0.3651
F-statistic:  97.8 on 3 and 502 DF,  p-value: < 2.2e-16

```

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

```

Call:
lm(formula = crim ~ poly(ptratio, 3))

Residuals:
    Min       1Q   Median       3Q      Max
-6.833 -4.146 -1.655  1.408 82.697

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.614      0.361  10.008 < 2e-16 ***
poly(ptratio, 3)1  56.045      8.122   6.901 1.57e-11 ***
poly(ptratio, 3)2  24.775      8.122   3.050 0.00241 **
poly(ptratio, 3)3 -22.280      8.122  -2.743 0.00630 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.122 on 502 degrees of freedom
Multiple R-squared:  0.1138,    Adjusted R-squared:  0.1085
F-statistic: 21.48 on 3 and 502 DF,  p-value: 4.171e-13

Call:
lm(formula = crim ~ poly(black, 3))

Residuals:
    Min       1Q   Median       3Q      Max
-13.096 -2.343 -2.128 -1.439 86.790

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.6135      0.3536  10.218 <2e-16 ***
poly(black, 3)1 -74.4312      7.9546  -9.357 <2e-16 ***
poly(black, 3)2   5.9264      7.9546   0.745  0.457
poly(black, 3)3  -4.8346      7.9546  -0.608  0.544
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.955 on 502 degrees of freedom
Multiple R-squared:  0.1498,    Adjusted R-squared:  0.1448
F-statistic: 29.49 on 3 and 502 DF,  p-value: < 2.2e-16

```

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

```

Call:
lm(formula = crim ~ poly(lstat, 3))

Residuals:
    Min       1Q   Median       3Q      Max
-15.234  -2.151  -0.486   0.066  83.353

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.6135     0.3392  10.654 <2e-16 ***
poly(lstat, 3)1  88.0697     7.6294  11.543 <2e-16 ***
poly(lstat, 3)2  15.8882     7.6294   2.082  0.0378 *
poly(lstat, 3)3 -11.5740     7.6294  -1.517  0.1299
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.629 on 502 degrees of freedom
Multiple R-squared:  0.2179,    Adjusted R-squared:  0.2133
F-statistic: 46.63 on 3 and 502 DF,  p-value: < 2.2e-16

Call:
lm(formula = crim ~ poly(medv, 3))

Residuals:
    Min       1Q   Median       3Q      Max
-24.427  -1.976  -0.437   0.439  73.655

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.614     0.292  12.374 < 2e-16 ***
poly(medv, 3)1  -75.058     6.569  -11.426 < 2e-16 ***
poly(medv, 3)2   88.086     6.569  13.409 < 2e-16 ***
poly(medv, 3)3  -48.033     6.569   -7.312 1.05e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.569 on 502 degrees of freedom
Multiple R-squared:  0.4202,    Adjusted R-squared:  0.4167
F-statistic: 121.3 on 3 and 502 DF,  p-value: < 2.2e-16

```

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.