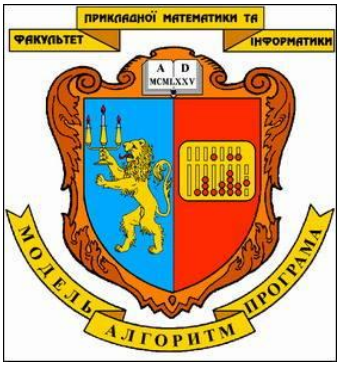
МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЛЬВІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ІВАНА ФРАНКА
Факультет прикладної математики та інформатики
Кафедра програмування



Індивідуальне завдання № 2

Лінійна регресія

Виконала:
студентка групи ПМОм-11
Кравець Ольга

Львів 2025

Хід роботи

Варіант - 3

Встановлюю значення змінної variant: для цього 0 (номер групи ПМОм-11)*25 + 3 (порядковий номер в списку групи) = 3.

```
> variant=3
> variant
[1] 3
```

Далі встановлюю set.seed(variant) та генерую значення змінної redundant як заокруглене до цілого випадкове число з рівномірного на інтервалі (5, 25) розподілу.

```
> set.seed(variant)
> redundant=floor(runif(1,5,25))
> redundant
[1] 8
```

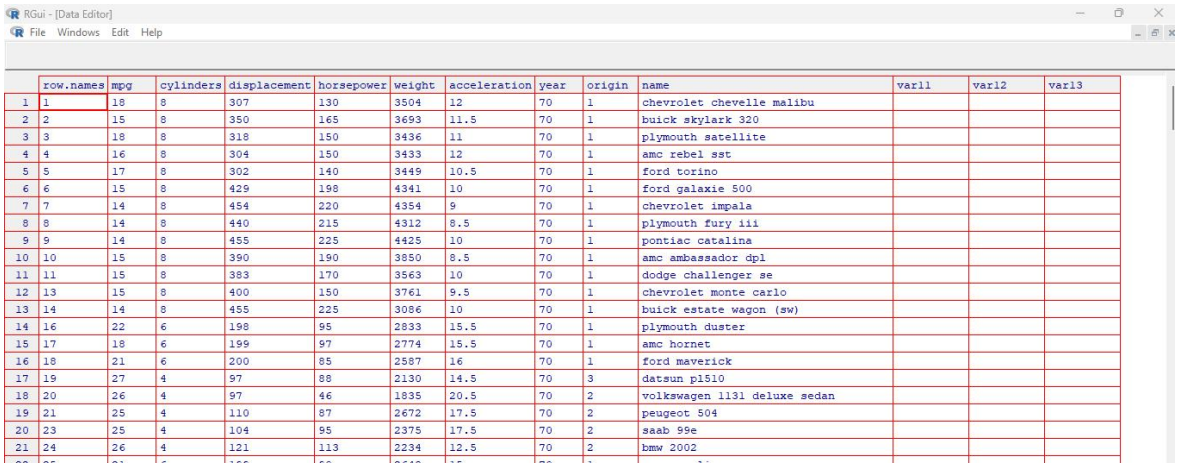
Зчитую дані з файлу “Auto.csv”

```
> setwd('D:\\Навчання\\Магістратура\\Моделі статистичного навчання\\02')
> Auto=read.csv('Auto.csv')
> names(Auto)
[1] "mpg"          "cylinders"    "displacement" "horsepower"   "weight"
[6] "acceleration" "year"         "origin"       "name"
```

Завдання 1.

Використовую функцію sample() для модифікації завантажених даних Auto - видалення redundant (% спостережень).

```
> Auto_new=Auto[-sample(1:length(Auto[,1]),round((redundant/100)*length(Auto[,1]))),]
> fix(Auto_new)
```



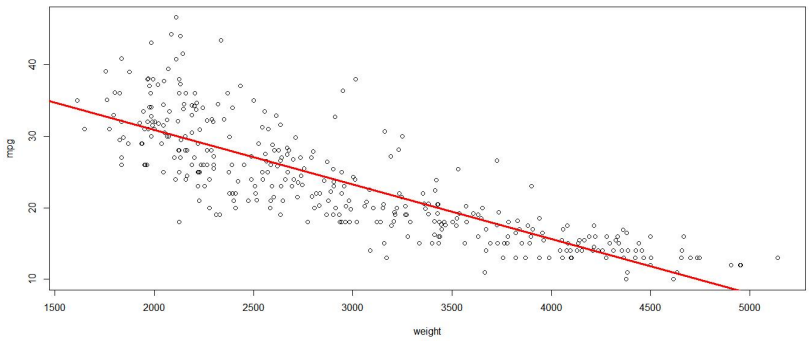
	row.names	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin	name	var11	var12	var13
1	1	18	8	307	130	3504	12	70	1	chevrolet chevelle malibu			
2	2	15	8	350	165	3693	11.5	70	1	buick skylark 320			
3	3	18	8	318	150	3436	11	70	1	plymouth satellite			
4	4	16	8	304	150	3433	12	70	1	amc rebel sst			
5	5	17	8	302	140	3449	10.5	70	1	ford torino			
6	6	15	8	429	198	4341	10	70	1	ford galaxie 500			
7	7	14	8	454	220	4354	9	70	1	chevrolet impala			
8	8	14	8	440	215	4312	8.5	70	1	plymouth fury iii			
9	9	14	8	455	225	4425	10	70	1	pontiac catalina			
10	10	15	8	390	190	3850	8.5	70	1	amc ambassador dpl			
11	11	15	8	383	170	3563	10	70	1	dodge challenger se			
12	13	15	8	400	150	3761	9.5	70	1	chevrolet monte carlo			
13	14	14	8	455	225	3086	10	70	1	buick estate wagon (sw)			
14	16	22	6	198	95	2833	15.5	70	1	plymouth duster			
15	17	18	6	199	97	2774	15.5	70	1	amc hornet			
16	18	21	6	200	85	2587	16	70	1	ford maverick			
17	19	27	4	97	88	2130	14.5	70	3	datsun pl510			
18	20	26	4	97	46	1835	20.5	70	2	volkswagen 1131 deluxe sedan			
19	21	25	4	110	87	2672	17.5	70	2	peugeot 504			
20	23	25	4	104	95	2375	17.5	70	2	saab 99e			
21	24	26	4	121	113	2234	12.5	70	2	bmw 2002			

Будую просту лінійну регресію з залежною змінною mpg і незалежною - weight, використовую функцію lm() для оцінки лінійної регресії та функцію summary() для виводу результату

```
> lrl=lm(mpg~weight,data=Auto_new)
> attach(Auto_new)
> summary(lrl)

Call:
lm(formula = mpg ~ weight, data = Auto_new)

Residuals:
     Min       1Q   Median       3Q      Max
-11.9599  -2.7699  -0.3738   2.1738  16.5332

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 46.1728343  0.8365430   55.20   <2e-16 ***
weight      -0.0076332  0.0002722  -28.04   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.383 on 363 degrees of freedom
Multiple R-squared:  0.6841,    Adjusted R-squared:  0.6833
F-statistic: 786.2 on 1 and 363 DF,  p-value: < 2.2e-16
```

Зобразила графічно предиктор та залежну змінну.

```
> plot(weight,mpg)
> abline(lrl,col='red',lwd=3)
```



Розкиданість точок навколо регресійної лінії змінюється. На лівій частині графіка точки більш розкидані, а праворуч вони стають щільнішими. Це може свідчити про зміну дисперсії залишків.

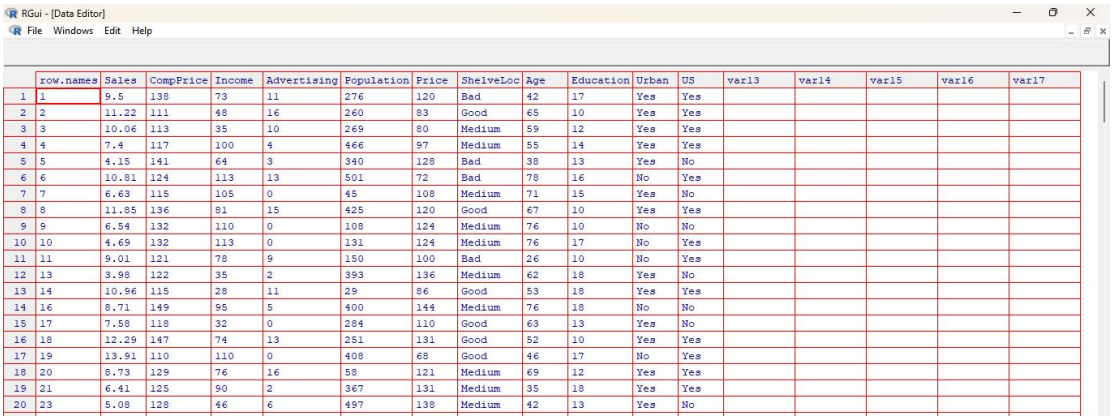
Графік підтверджує наявність негативного лінійного зв’язку між weight та mpg, тому weight - росте, а mpg - знижується.

Значення p-значення для weight свідчить про значущий вплив ваги авто на витрати пального.

Завдання 2.

Модифікую дані Carseats, аналогічно до завдання 1.

```
> library(ISLR)
> names(Carseats)
 [1] "Sales"       "CompPrice"   "Income"      "Advertising" "Population"
 [6] "Price"       "ShelveLoc"   "Age"         "Education"   "Urban"
[11] "US"
> Car=Carseats[-sample(1:length(Carseats[,1]),round((redundant/100)*length(Carseats[,1]))),]
> fix(Car)
```



RGui - [Data Editor]

File Windows Edit Help

	row.names	Sales	CompPrice	Income	Advertising	Population	Price	ShelveLoc	Age	Education	Urban	US	var13	var14	var15	var16	var17
1	1	9.5	138	73	11	276	120	Bad	42	17	Yes	Yes					
2	2	11.22	111	48	16	260	83	Good	65	10	Yes	Yes					
3	3	10.06	113	35	10	269	80	Medium	59	12	Yes	Yes					
4	4	7.4	117	100	4	466	97	Medium	55	14	Yes	Yes					
5	5	4.15	141	64	3	340	128	Bad	38	13	Yes	No					
6	6	10.81	124	113	13	501	72	Bad	78	16	No	Yes					
7	7	6.63	115	105	0	45	108	Medium	71	15	Yes	No					
8	8	11.85	136	81	15	425	120	Good	67	10	Yes	Yes					
9	9	6.54	132	110	0	108	124	Medium	76	10	No	No					
10	10	4.69	132	113	0	131	124	Medium	76	17	No	Yes					
11	11	9.01	121	78	9	150	100	Bad	26	10	No	Yes					
12	13	3.98	122	35	2	393	136	Medium	62	18	Yes	No					
13	14	10.96	115	28	11	29	86	Good	53	18	Yes	Yes					
14	16	8.71	149	95	5	400	144	Medium	76	18	No	No					
15	17	7.58	118	32	0	284	110	Good	63	13	Yes	No					
16	18	12.29	147	74	13	251	131	Good	52	10	Yes	Yes					
17	19	13.91	110	110	0	408	68	Good	46	17	No	Yes					
18	20	8.73	129	76	16	58	121	Medium	69	12	Yes	Yes					
19	21	6.41	125	90	2	367	131	Medium	35	18	Yes	Yes					
20	23	5.08	128	46	6	497	138	Medium	42	13	Yes	No					

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

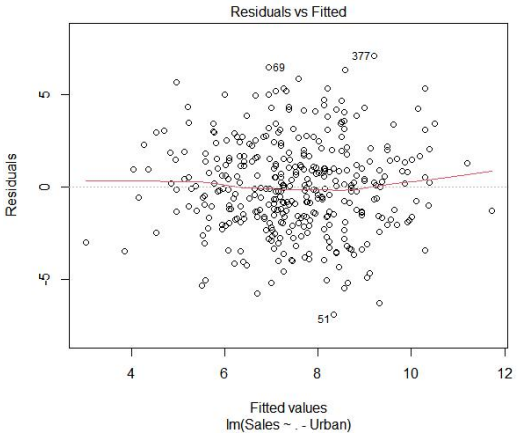
```
> Car_2=Car[,-c(2:5, 7:9)]
> fix(Car_2)
> lr_2=lm(Sales~.,data=Car_2)
> summary(lr_2)

Call:
lm(formula = Sales ~ ., data = Car_2)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9386 -1.6558 -0.0659  1.5193  7.0460

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.989273   0.668952  19.417  < 2e-16 ***
Price       -0.054085   0.005404 -10.009  < 2e-16 ***
UrbanYes     0.050996   0.286721   0.178    0.859
USYes        1.159523   0.271992   4.263 2.57e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.488 on 364 degrees of freedom
Multiple R-squared:  0.2372,    Adjusted R-squared:  0.2309
F-statistic: 37.73 on 3 and 364 DF,  p-value: < 2.2e-16
```

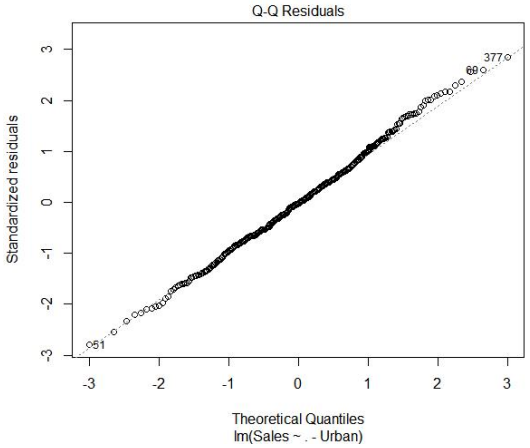
Яку зі змінних можна вилучити з моделі? Бачу, що Urban не має істотного зв'язку зі Sales (p-значення 0.859), тому будую нову модель вже без Urban.

```
> lr_3=lm(Sales~.-Urban,data=Car_2)
> summary(lr_3)

Call:
lm(formula = Sales ~ . - Urban, data = Car_2)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9243 -1.6463 -0.0688  1.5260  7.0614

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.016538   0.650285  20.017  < 2e-16 ***
Price       -0.054024   0.005385 -10.032  < 2e-16 ***
USYes        1.162271   0.271192   4.286 2.33e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.485 on 365 degrees of freedom
Multiple R-squared:  0.2371,    Adjusted R-squared:  0.233
F-statistic: 56.73 on 2 and 365 DF,  p-value: < 2.2e-16
```

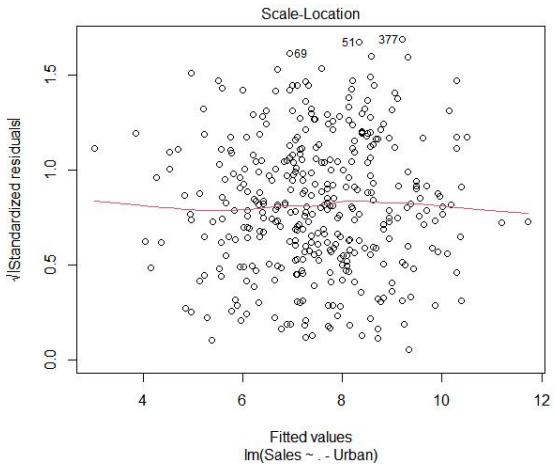
Далі будую діагностичні графіки моделі

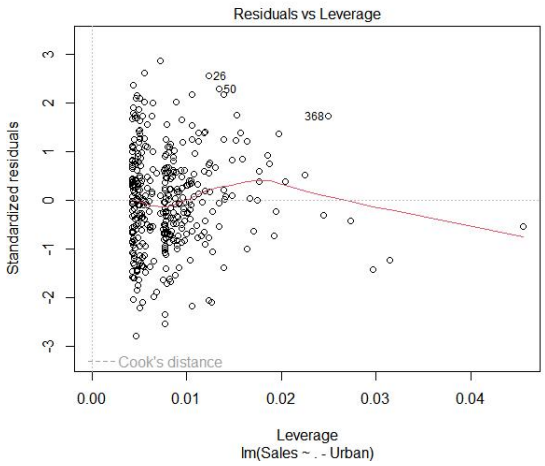
```
> plot(lr_3)
Waiting to confirm page change...
Waiting to confirm page change...
Waiting to confirm page change...
Waiting to confirm page change...
```





Є кілька точок, які виходять за межі норми (51, 69 та 377).





На графіку залишків від оцінених значень видно, що залежність не є лінійною. Також бачу, що є викиди і що спостереження мають високий рівень левереджу.

```
> length(Car_2$Price)
[1] 368
> H=hatvalues(lr_3)
> max(H)
[1] 0.04697393
> ResStud=rstudent(lr_3)
> max(ResStud)
[1] 2.885913
```

Найбільший стандартизований залишковий викид - 2.88.

Завдання 3.

Виконую команди вказані в завданні 3.

```
> set.seed(variant)
> x1=runif (100)
> x2 =(variant/2)* x1+rnorm (100) *variant/10
> y=(2*variant)+variant* x1 +(variant/3)* x2+rnorm (100)
```

Яка кореляція між x1 та x2?

```
> cor(x1,x2)
[1] 0.8425111
```

Кореляція між x1 та x2 є позитивною, що вказує на досить сильний лінійний зв'язок між змінними.

Оцінюю регресію методом найменших квадратів, щоб передбачити y, використовуючи x1 та x2.

```
> lr_4=lm(y~x1+x2)
> summary(lr_4)

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

Residuals:
     Min       1Q   Median       3Q      Max
-2.10475 -0.85724  0.06439  0.71709  2.53733

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.8167     0.2187  26.597  < 2e-16 ***
x1            4.1700     0.7227   5.770 9.46e-08 ***
x2            0.6716     0.4042   1.661   0.0999 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.109 on 97 degrees of freedom
Multiple R-squared:  0.6497,    Adjusted R-squared:  0.6425
F-statistic: 89.96 on 2 and 97 DF,  p-value: < 2.2e-16
```

Чи можна відхилити нульову гіпотезу H0: $\beta 1 = 0$?

p-значення x1 < 0.05. Гіпотезу відхиляємо.

Як щодо гіпотези H0: $\beta 2 = 0$?

p-значення x2 > 0.05. Гіпотезу приймаємо.

Будую регресію y на x1.

```
> lr_5=lm(y~x1)
> summary(lr_5)

Call:
lm(formula = y ~ x1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.08958 -0.89613 -0.04303  0.78733  2.46323

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.8103     0.2206   26.34   <2e-16 ***
x1            5.1816     0.3928   13.19   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.119 on 98 degrees of freedom
Multiple R-squared:  0.6398,    Adjusted R-squared:  0.6361
F-statistic:   174 on 1 and 98 DF,  p-value: < 2.2e-16
```

Чи можна відхилити нульову гіпотезу H0: β1 = 0?

p-значення x1 < 0.05. Гіпотезу відхиляємо.

Будую регресію y на x2.

```
> lr_6=lm(y~x2)
> summary(lr_6)

Call:
lm(formula = y ~ x2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.01561 -1.03205 -0.07644  0.97823  3.05114

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.4212     0.2214   29.01   <2e-16 ***
x2            2.6369     0.2511   10.50   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.279 on 98 degrees of freedom
Multiple R-squared:  0.5295,    Adjusted R-squared:  0.5247
F-statistic: 110.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

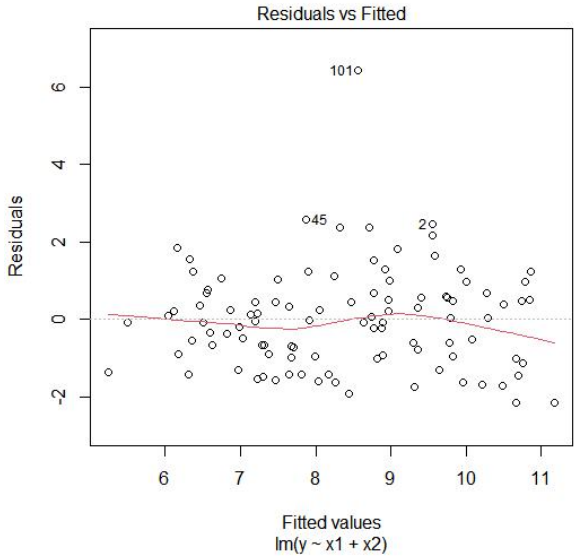
Чи можна відхилити нульову гіпотезу H0: β2 = 0?

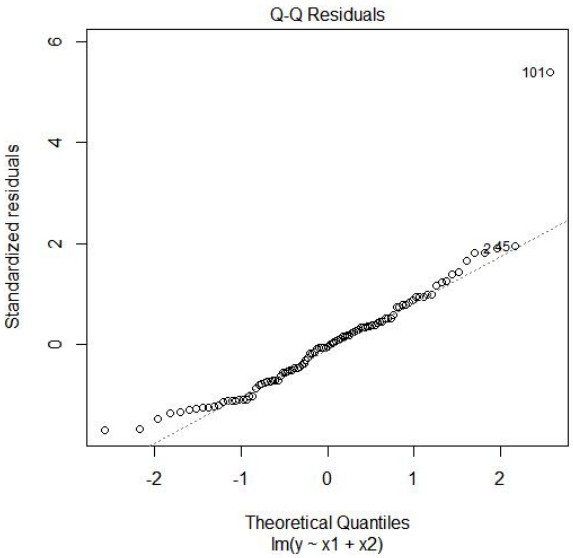
p-значення x1 < 0.05. Гіпотезу відхиляємо.

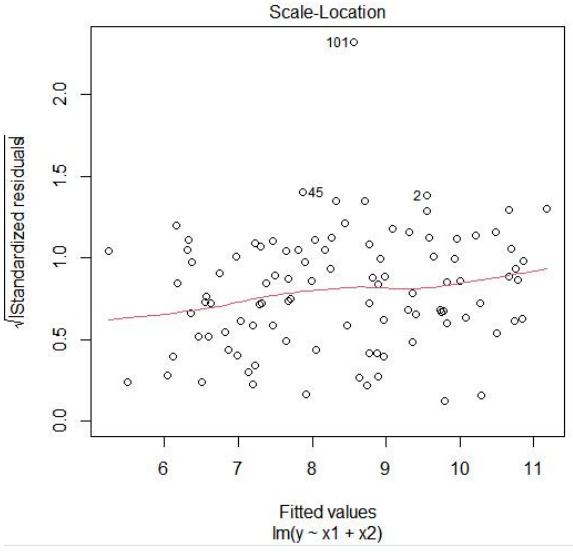
Нехай одне додаткове спостереження було неправильно виміряно. Переоцінюю попередні лінійні моделі, використовуючи ці нові дані.

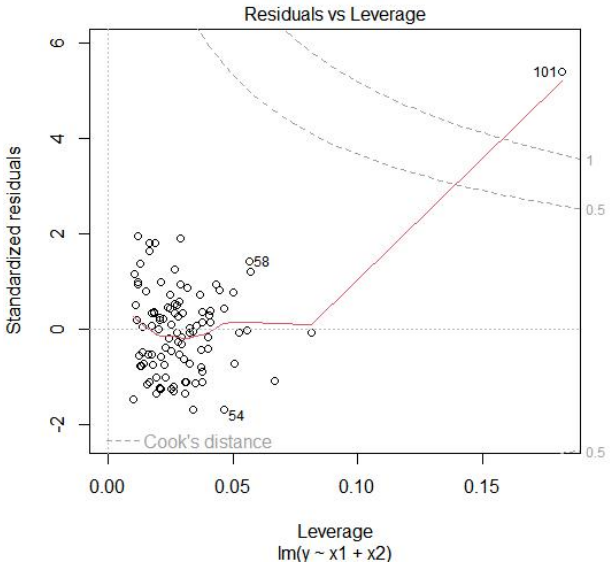
```
> x1 = c (x1, 0.1)
> x2 = c (x2, (variant/2)* 0.9)
> y = c (y, 5*variant)
> lr_7=lm(y~x1+x2)
> summary(lr_7)

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

Residuals:
    Min      1Q  Median      3Q     Max
-2.1725 -0.9501 -0.0350  0.6764  6.4301

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.0381     0.2563  23.555  < 2e-16 ***
x1            2.3105     0.7857   2.941 0.004084 **
x2            1.7042     0.4401   3.872 0.000194 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.317 on 98 degrees of freedom
Multiple R-squared:  0.5586,    Adjusted R-squared:  0.5496
F-statistic: 62.02 on 2 and 98 DF,  p-value: < 2.2e-16
```

p-значення x1 < 0.05, p-значення x2 < 0.05.









Є кілька точок, які виходять за межі норми (2, 45) і точка, яка дуже віддалена (101).

Досліджую отриману модель на наявність викидів та спостережень з високим рівнем левереджу

```
> H7=hatvalues(lr_7)
> max(H7)
[1] 0.1818779
> length(x1)
[1] 101
> ResStud7=rstudent(lr_7)
> max(ResStud7)
[1] 6.40947
```

Найбільший стандартизований залишковий викид - 6.4.