

Кореляційний та регресійний аналіз

Продовжимо працювати з лінійною регресією та трактуванням результатів отриманих моделей.

Будемо використовувати два набори даних: квартет Анскомбе `anscombe` (згенерований у 1973 Frensis Anscombe) та `diamonds` (містить інформацію про ціну та характеристики 53940 діамантів).

Почнемо з дослідження набору даних `anscombe`. На цьому наборі даних проілюструємо процес діагностики моделі лінійної регресії. Це вбудований R датасет, тому завантажувати його додатково не потрібно. Дослідимо структуру наших даних:

```
anscombe
```

	x1	x2	x3	x4	y1	y2	y3	y4
1	10	10	10	8	8.04	9.14	7.46	6.58
2	8	8	8	8	6.95	8.14	6.77	5.76
3	13	13	13	8	7.58	8.74	12.74	7.71
4	9	9	9	8	8.81	8.77	7.11	8.84
5	11	11	11	8	8.33	9.26	7.81	8.47
6	14	14	14	8	9.96	8.10	8.84	7.04
7	6	6	6	8	7.24	6.13	6.08	5.25
8	4	4	4	19	4.26	3.10	5.39	12.50
9	12	12	12	8	10.84	9.13	8.15	5.56
10	7	7	7	8	4.82	7.26	6.42	7.91
11	5	5	5	8	5.68	4.74	5.73	6.89

```
str(anscombe)
```

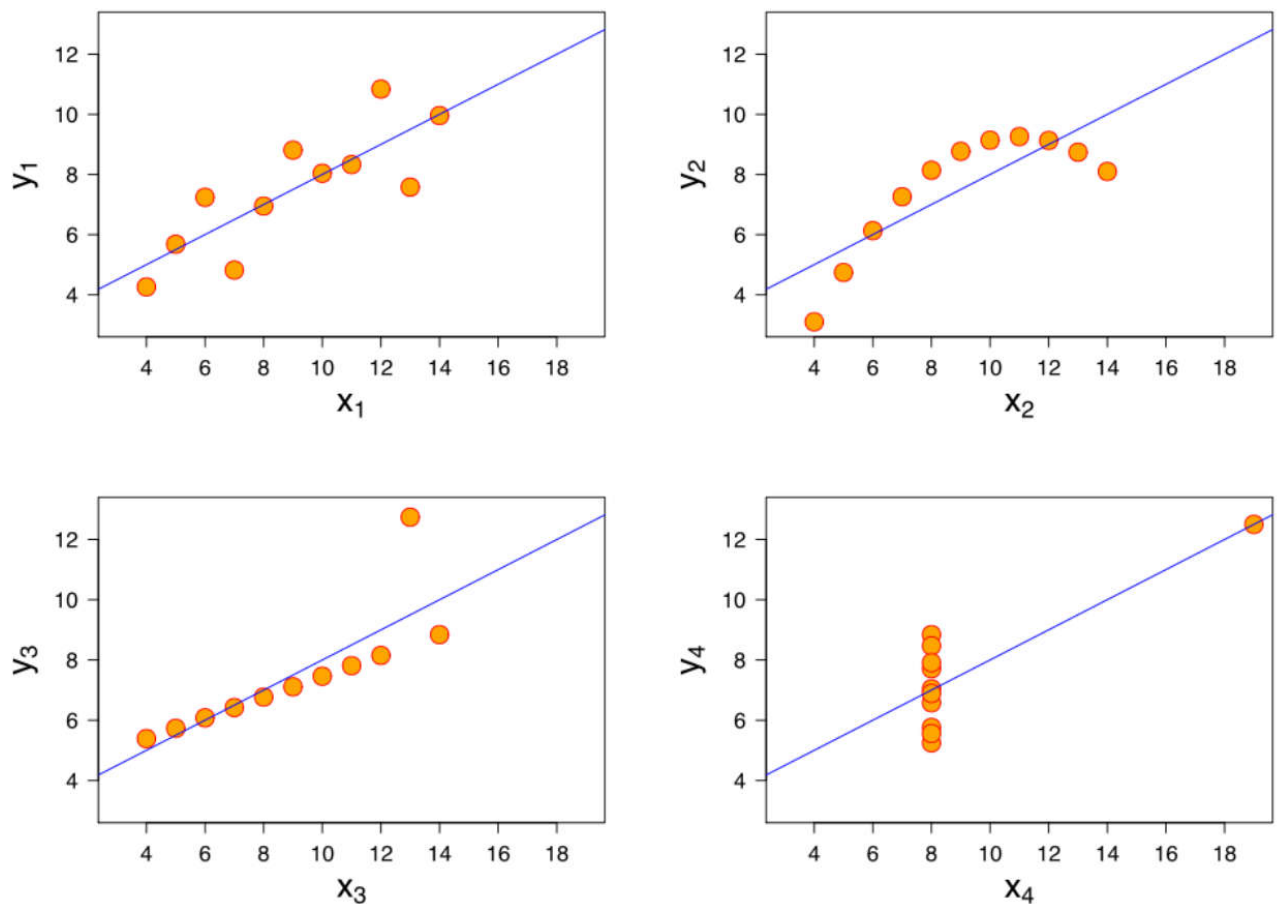
```
'data.frame':  11 obs. of  8 variables:
 $ x1: num  10 8 13 9 11 14 6 4 12 7 ...
 $ x2: num  10 8 13 9 11 14 6 4 12 7 ...
 $ x3: num  10 8 13 9 11 14 6 4 12 7 ...
 $ x4: num   8 8 8 8 8 8 8 19 8 8 ...
 $ y1: num  8.04 6.95 7.58 8.81 8.33 ...
 $ y2: num  9.14 8.14 8.74 8.77 9.26 8.1 6.13 3.1 9.13 7.26 ...
 $ y3: num  7.46 6.77 12.74 7.11 7.81 ...
 $ y4: num  6.58 5.76 7.71 8.84 8.47 7.04 5.25 12.5 5.56 7.91 ...
```

```
summary(anscombe)
```

x1		x2		x3		x4	
Min.	: 4.0	Min.	: 4.0	Min.	: 4.0	Min.	: 8
1st Qu.	: 6.5	1st Qu.	: 6.5	1st Qu.	: 6.5	1st Qu.	: 8
Median	: 9.0	Median	: 9.0	Median	: 9.0	Median	: 8
Mean	: 9.0	Mean	: 9.0	Mean	: 9.0	Mean	: 9
3rd Qu.	:11.5	3rd Qu.	:11.5	3rd Qu.	:11.5	3rd Qu.	: 8
Max.	:14.0	Max.	:14.0	Max.	:14.0	Max.	:19

y1		y2		y3		y4	
Min.	: 4.260	Min.	:3.100	Min.	: 5.39	Min.	: 5.250
1st Qu.	: 6.315	1st Qu.	:6.695	1st Qu.	: 6.25	1st Qu.	: 6.170
Median	: 7.580	Median	:8.140	Median	: 7.11	Median	: 7.040
Mean	: 7.501	Mean	:7.501	Mean	: 7.50	Mean	: 7.501
3rd Qu.	: 8.570	3rd Qu.	:8.950	3rd Qu.	: 7.98	3rd Qu.	: 8.190
Max.	:10.840	Max.	:9.260	Max.	:12.74	Max.	:12.500

Це чотири набори даних (x1, y1), (x2, y2), (x3, y3), (x4, y4)



Ці набори мають однаковий коефіцієнт кореляції (обчисліть його та вкажіть у якості відповіді на питання 1) та однакову лінію моделі лінійної регресії (знайдіть лінію регресії з допомогою команд **lm** та **summary**; вкажіть рівняння цієї лінії у якості відповіді на питання 2).

Умовами для побудови валідної моделі є:

- Лінійність
- Нормальний розподіл залишків
- Гомоскедастичність (стала варіативність залишків)

Проведемо діагностику лінійних моделей для (x_1, y_1) , (x_2, y_2) , (x_3, y_3) , (x_4, y_4) відповідно. З графіка очевидно, що для наборів (x_2, y_2) та (x_4, y_4) порушена умова лінійності. У випадку (x_2, y_2) – є нелінійна залежність. Для набору (x_3, y_3) умова лінійності буде виконуватись при видаленні нетипових значень (outliers), при цьому зміниться рівняння лінії лінійної регресії.

Аналіз залишків будемо проводити для всіх моделей (з метою зрозуміти, як буде виглядати розподіл та варіативність залишків при порушеннях умов лінійності).

Почнемо з набору (x_1, y_1) .

Нехай модель задана рівнянням `lm1 <- lm(data = anscombe, y1 ~ x1)`. Для діагностики моделі нам потрібно оцінити розподіл залишків.

Залишок - це різниця між реальними даними (в нашому випадку це y_1) та даними \hat{y} , для x_1 згідно моделі.

Знайдемо значення \hat{y} (fitted.values) згідно нашого рівняння лінійної регресії.

```
lm1$fitted.values
```

```
      1      2      3      4      5      6      7      8      9     10     11
8.001000 7.000818 9.501273 7.500909 8.501091 10.001364 6.000636 5.000455 9.001182 6.500727 5.500545
```

Знайдемо залишки. Це можна зробити віднявши fitted.values від реальних значень y_1 :

```
anscombe$y1 - lm1$fitted.values
```

```
      1      2      3      4      5      6      7      8      9
0.03900000 -0.05081818 -1.92127273 1.30909091 -0.17109091 -0.04136364 1.23936364 -0.74045455 1.83881818
     10     11
-1.68072727 0.17945455
```

або використати параметр **residuals**

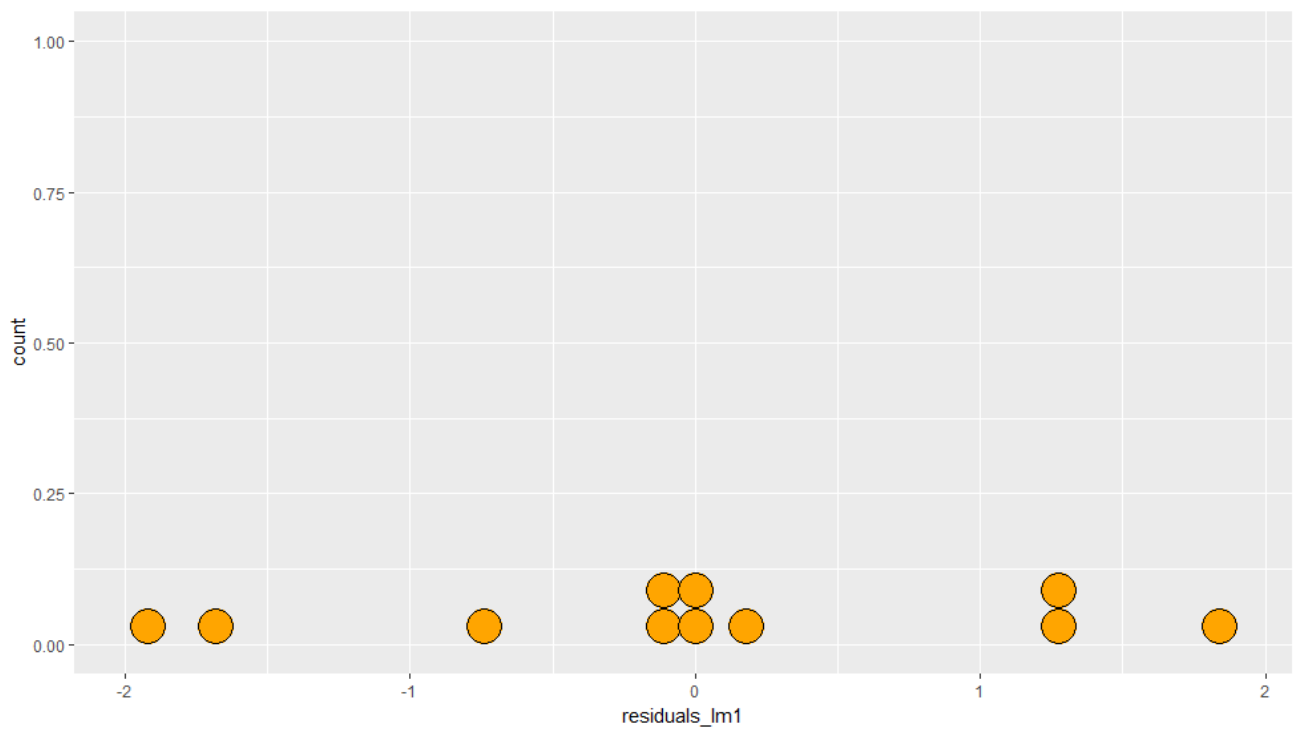
```
lm1$residuals
```

```
      1      2      3      4      5      6      7      8      9  
0.03900000 -0.05081818 -1.92127273  1.30909091 -0.17109091 -0.04136364  1.23936364 -0.74045455  1.83881818  
     10     11  
-1.68072727  0.17945455
```

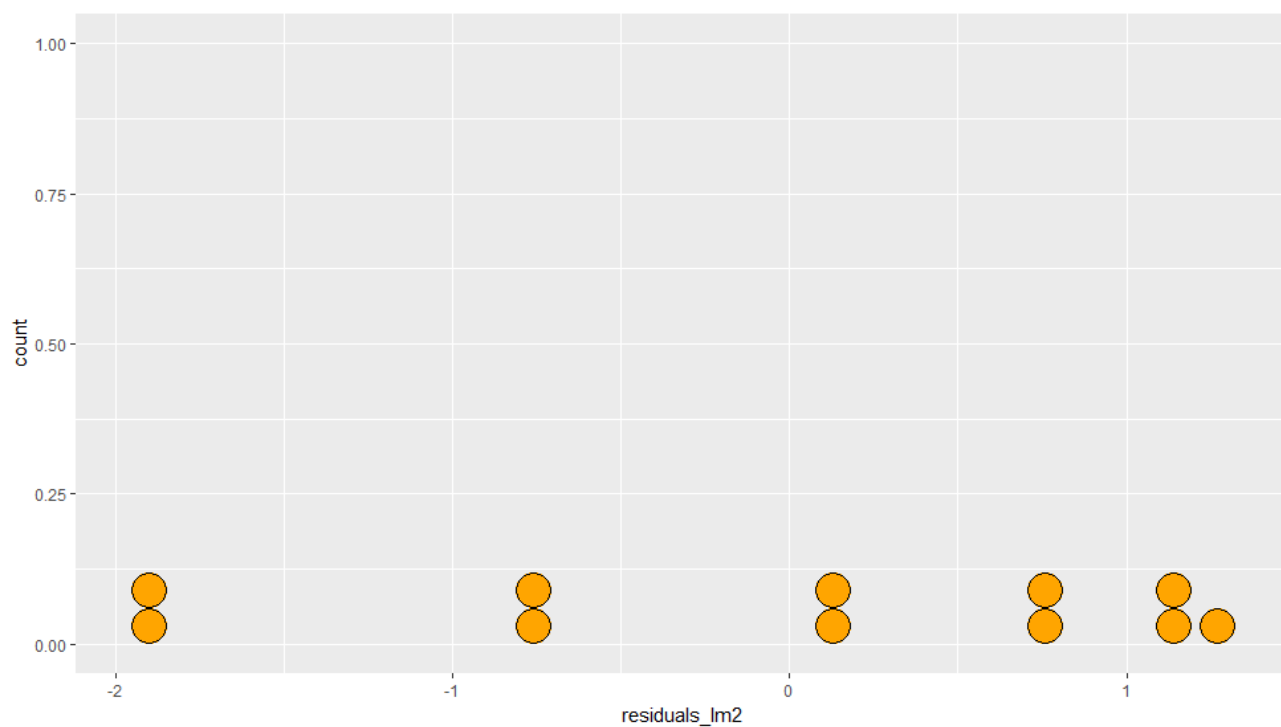
Найкраще оцінювати розподіл даних з допомогою гістограми, але так як у нас всього одинадцять точок, можемо використати точковий графік.

```
library(ggplot2)
```

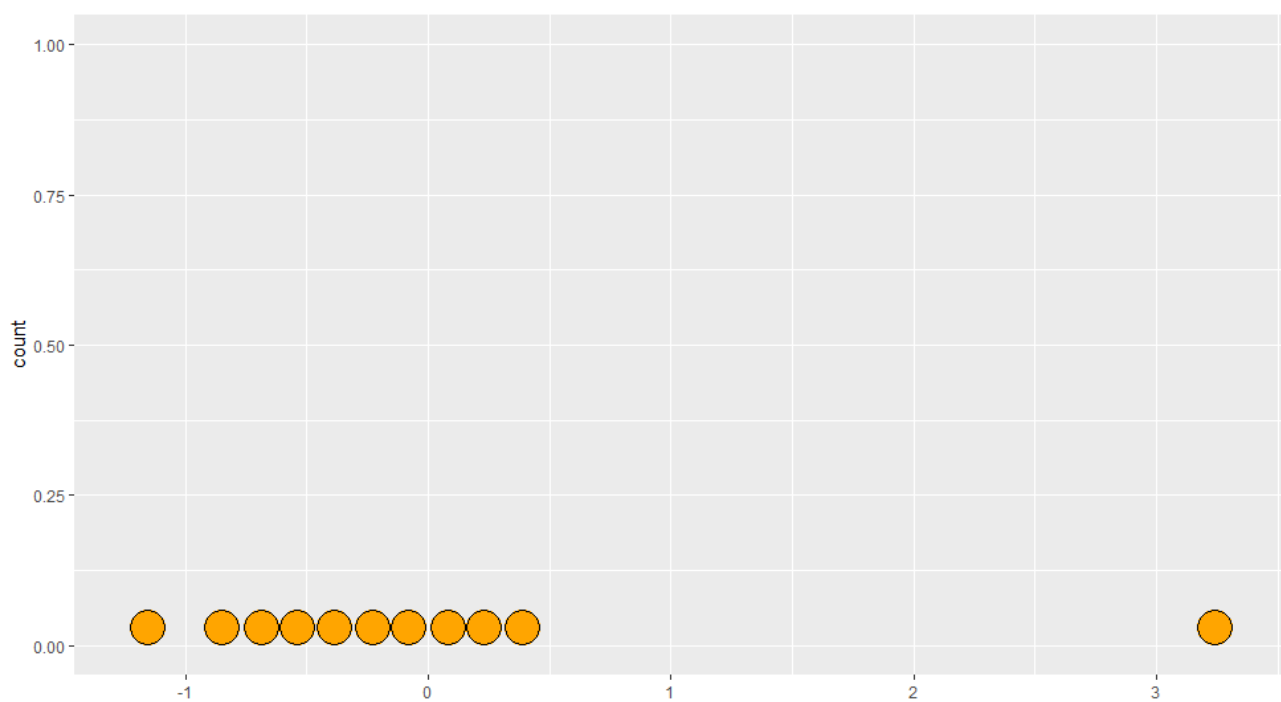
```
anscombe$residuals_lm1 <- lm1$residuals  
ggplot(anscombe, aes(x = residuals_lm1)) +  
  geom_dotplot(fill = "orange")
```



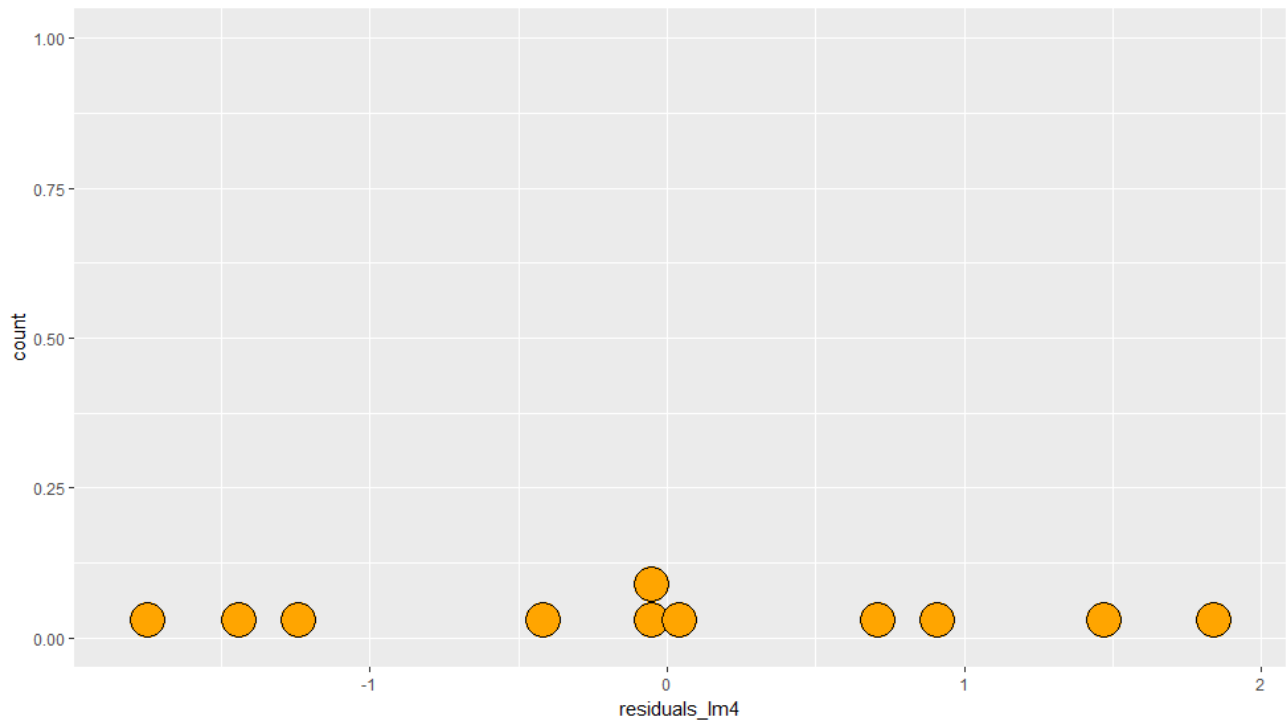
```
anscombe$residuals_lm2 <- lm2$residuals
ggplot(anscombe, aes(x = residuals_lm2)) +
  geom_dotplot(fill = "orange")
```



```
anscombe$residuals_lm3 <- lm3$residuals
ggplot(anscombe, aes(x = residuals_lm3)) +
  geom_dotplot(fill = "orange")
```



```
anscombe$residuals_lm4 <- lm4$residuals
ggplot(anscombe, aes(x = residuals_lm4)) +
  geom_dotplot(fill = "orange")
```



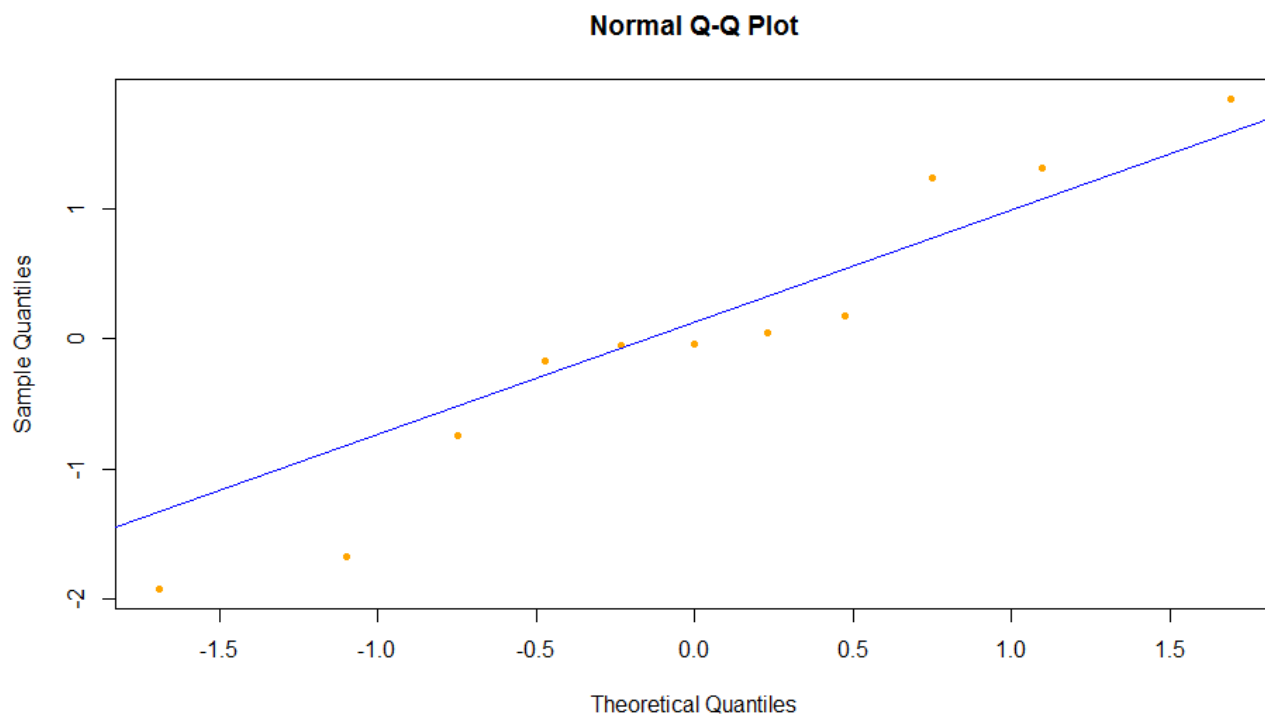
Оцінювати візуально розподіл даних для 11 точок досить важко, найбільше відповідають нормальному розподілу перший та третій набори даних.

Для оцінки нормальності розподілу, будемо використовувати функції **qqnorm** та **qqline**. Бібліотека **ggplot2** для цього аналізу менш зручна. Графіки для наших наборів будуть виглядати так:

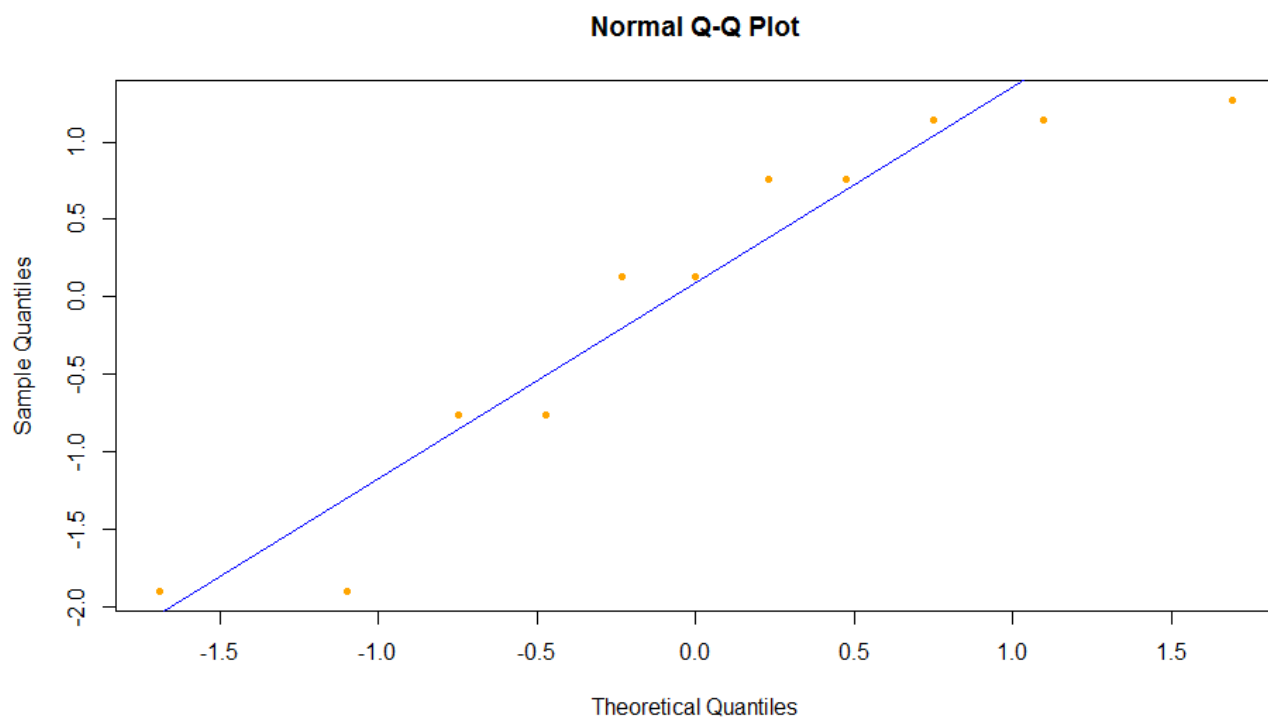
Для (x1, y1):

```
qqnorm(lm1$residuals, col="orange", pch=20)
```

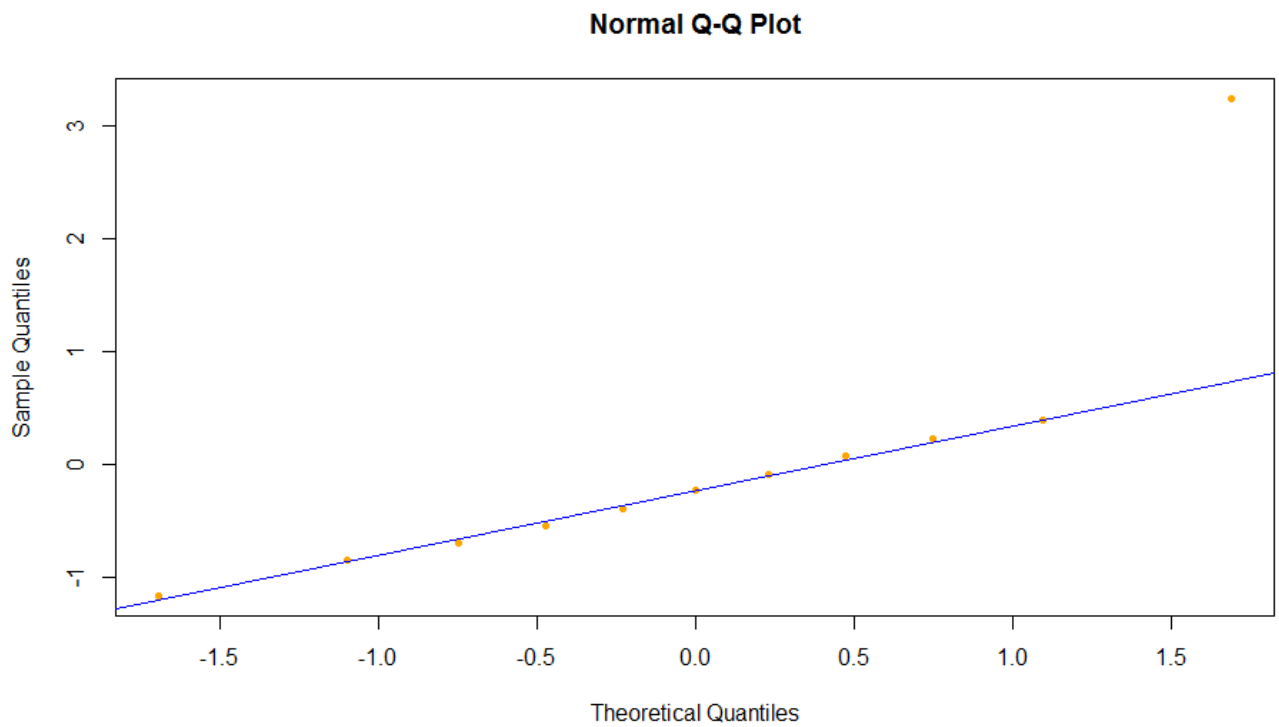
```
qqline(lm1$residuals, col = "blue")
```



Для (x2, y2):

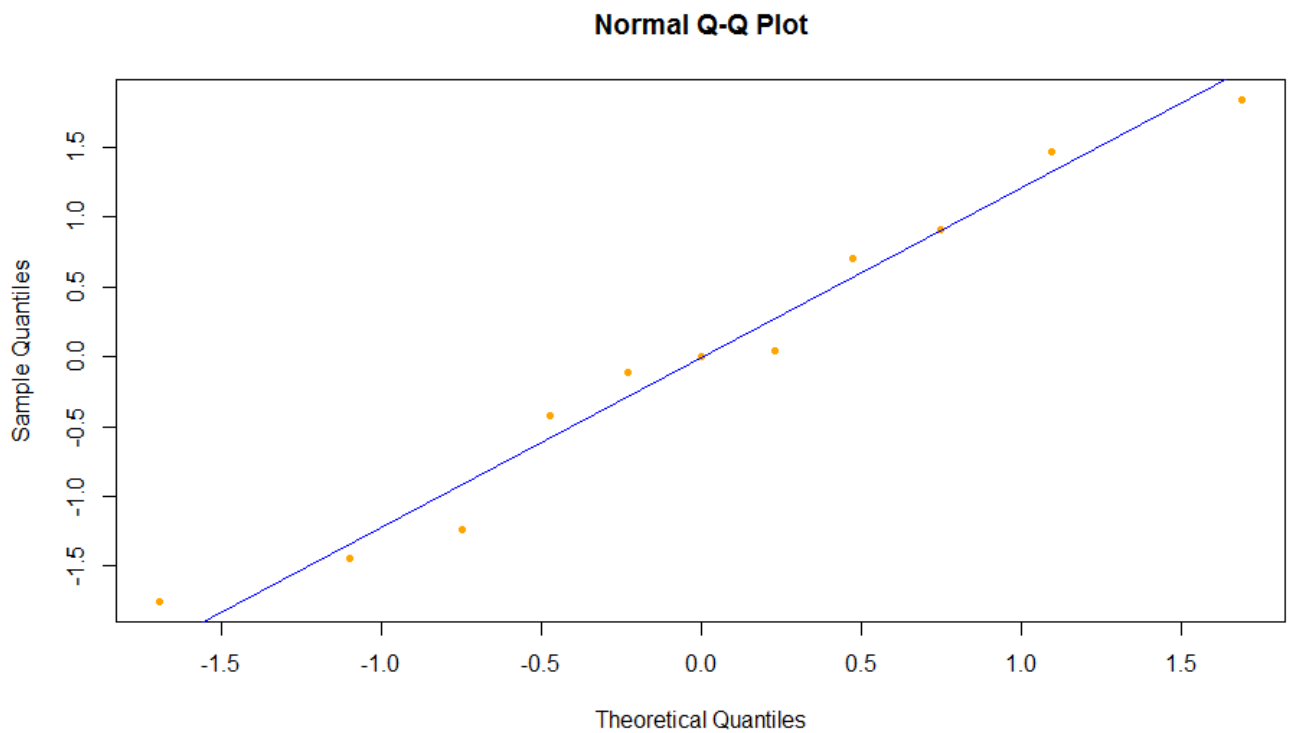


Для (x3, y3):



Для (x4, y4):

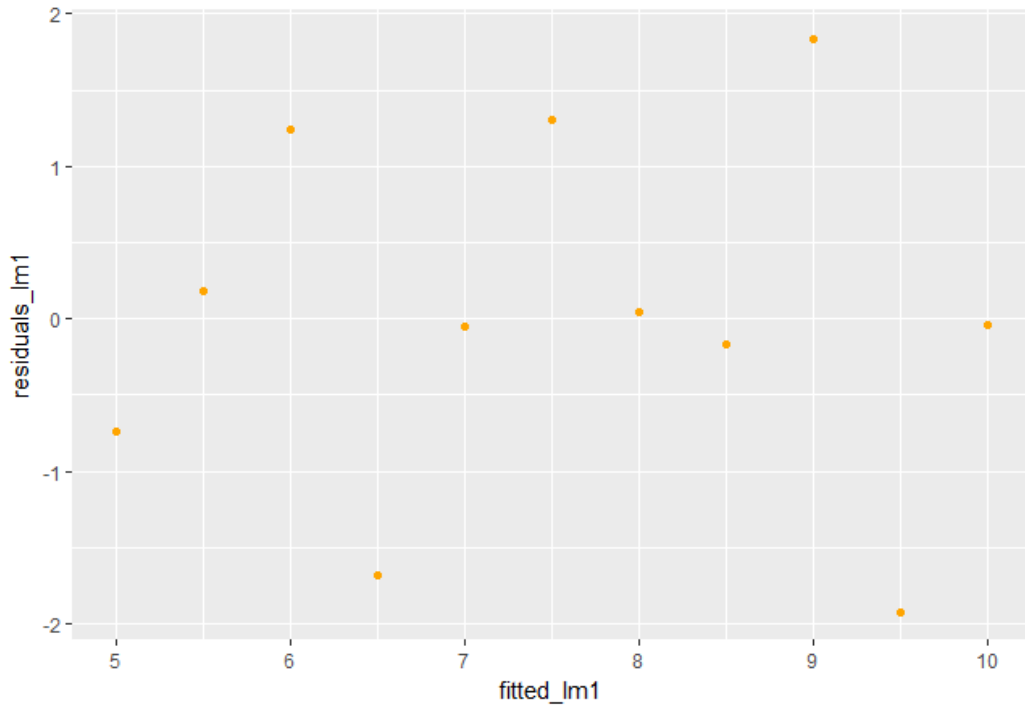
```
qqnorm(lm1$residuals, col="orange", pch=20)  
qqline(lm1$residuals, col = "blue")
```



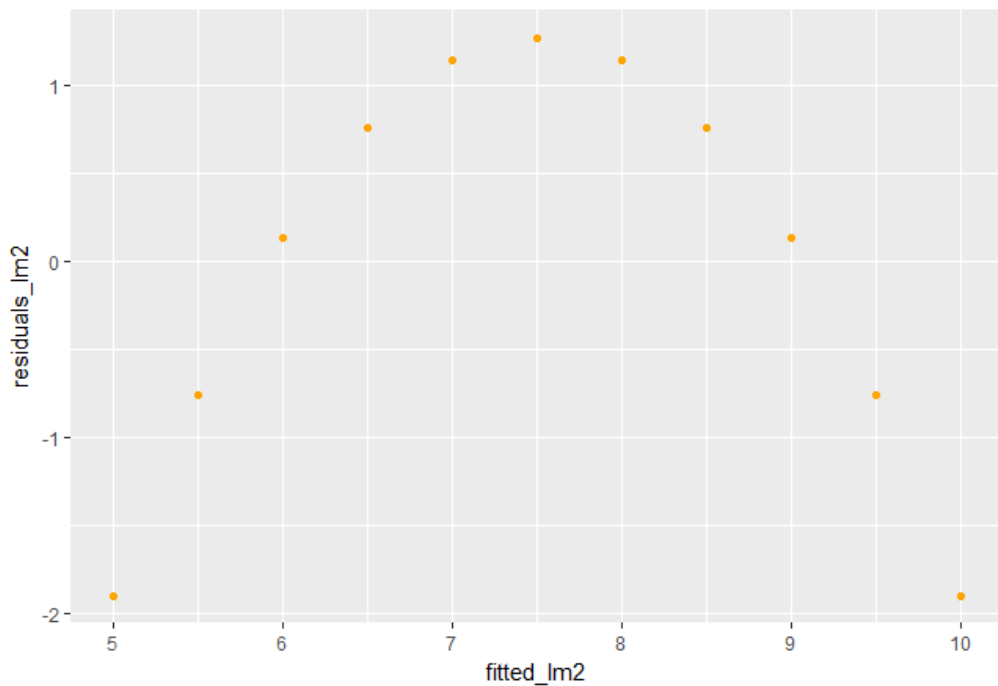
Оцінюємо варіативність залишків:

Для (x_1, y_1) умова сталості залишків виконується.

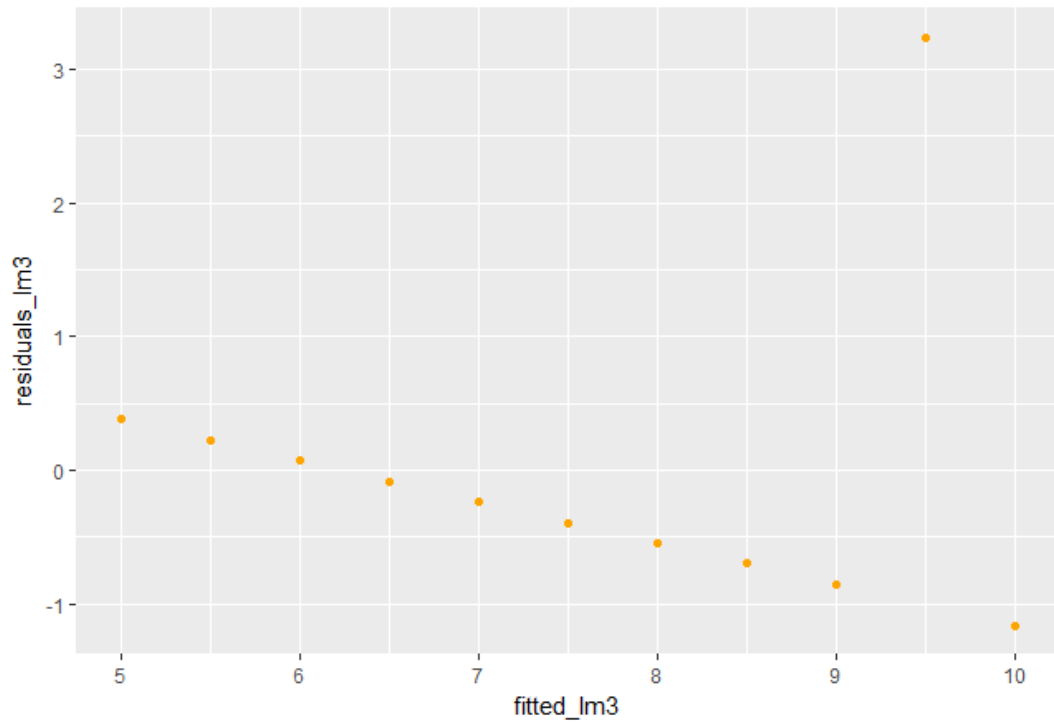
```
anscombe$fitted_lm1 <- lm1$fitted.values  
ggplot(data=anscombe, aes(x=fitted_lm1, y=residuals_lm1)) +  
  geom_point(col="orange")
```



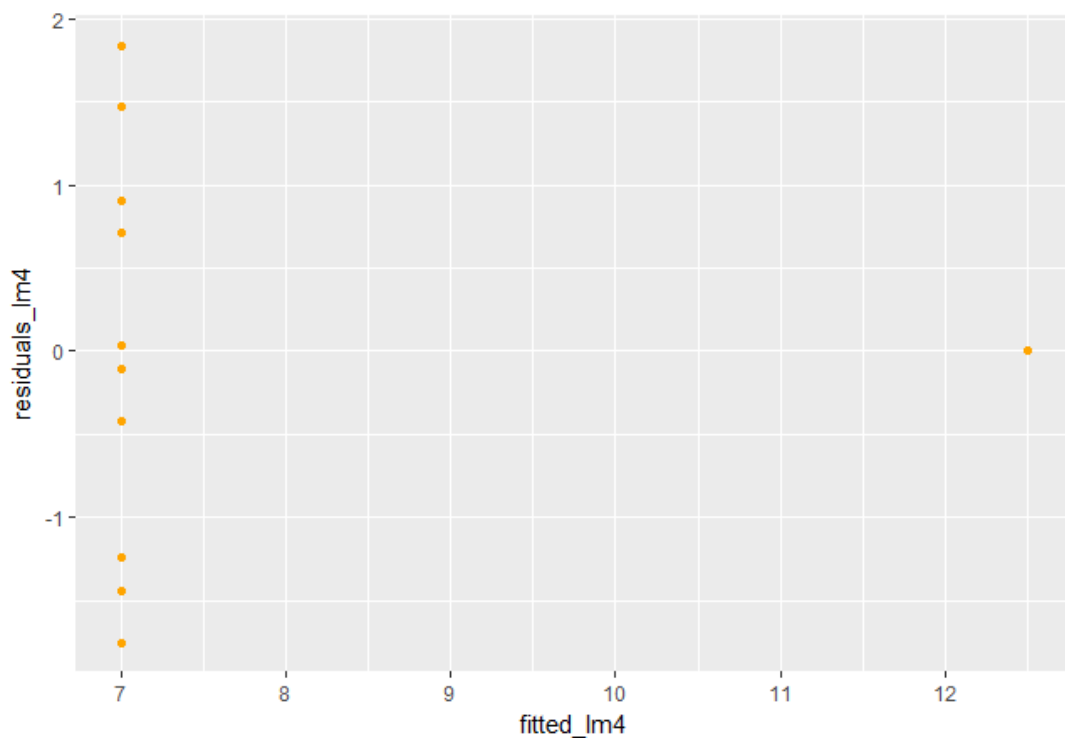
Для (x_2, y_2) умова сталості залишків не виконується.



Для (x_3, y_3) умова сталості залишків не виконується.



Для (x_4, y_4) умова сталості залишків не виконується.



На основі проведених досліджень, можемо стверджувати, що умови для побудови моделі лінійної залежності виконуються лише для набору (x_1, y_1) .

Перейдемо до реального набору даних diamonds, який містить інформацію про ціну та характеристики 53940 діамантів. Це вбудований набір даних бібліотеки ggplot2. За цим посиланням

http://varianceexplained.org/RData/code/code_lesson2/

ви можете ознайомитись з прикладами візуального аналізу цього набору даних.

В лабораторній ми будемо досліджувати залежність ціни від ваги діамантів.

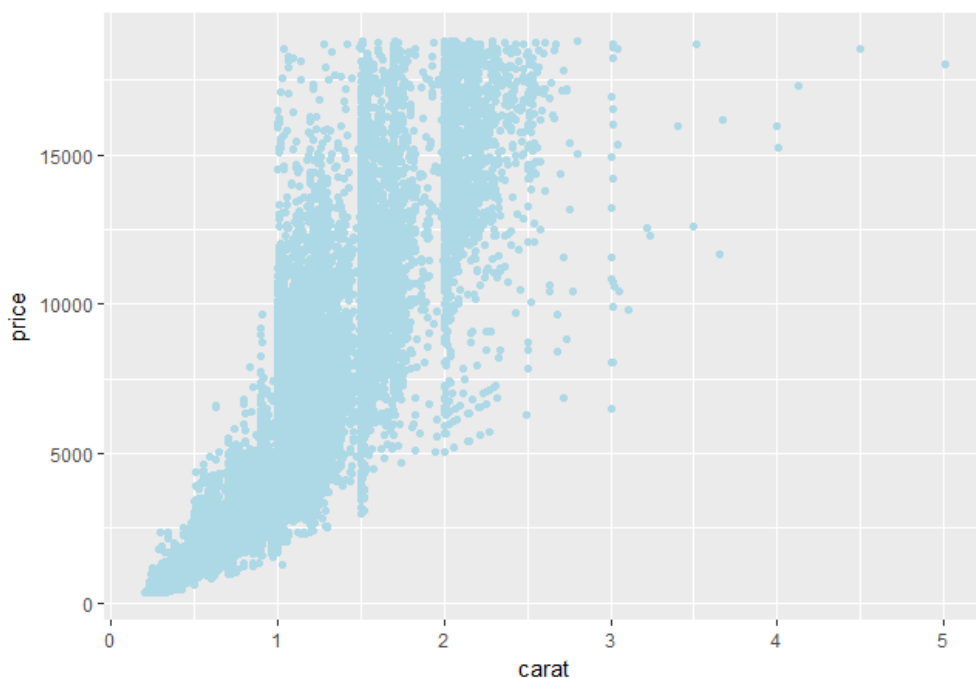
Подивимось на структуру набору даних:

```
str(diamonds)
Classes 'tbl_df', 'tbl' and 'data.frame':    53940 obs. of  10 variables:
 $ carat   : num  0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
 $ cut     : Ord.factor w/ 5 levels "Fair"<"Good"<...: 5 4 2 4 2 3 3 3 1 3 ...
 $ color   : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<...: 2 2 2 6 7 7 6 5 2 5 ...
 $ clarity : Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<...: 2 3 5 4 2 6 7 3 4 5 ...
 $ depth   : num  61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
 $ table   : num  55 61 65 58 58 57 57 55 61 61 ...
 $ price   : int  326 326 327 334 335 336 336 337 337 338 ...
 $ x       : num  3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
 $ y       : num  3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
 $ z       : num  2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
```

В нас є 10 змінних та 53940 спостережень. Будемо досліджувати залежність між вагою (змінна **carat**) та ціною (змінна **price**).

Побудуємо графік розсіювання для цих змінних:

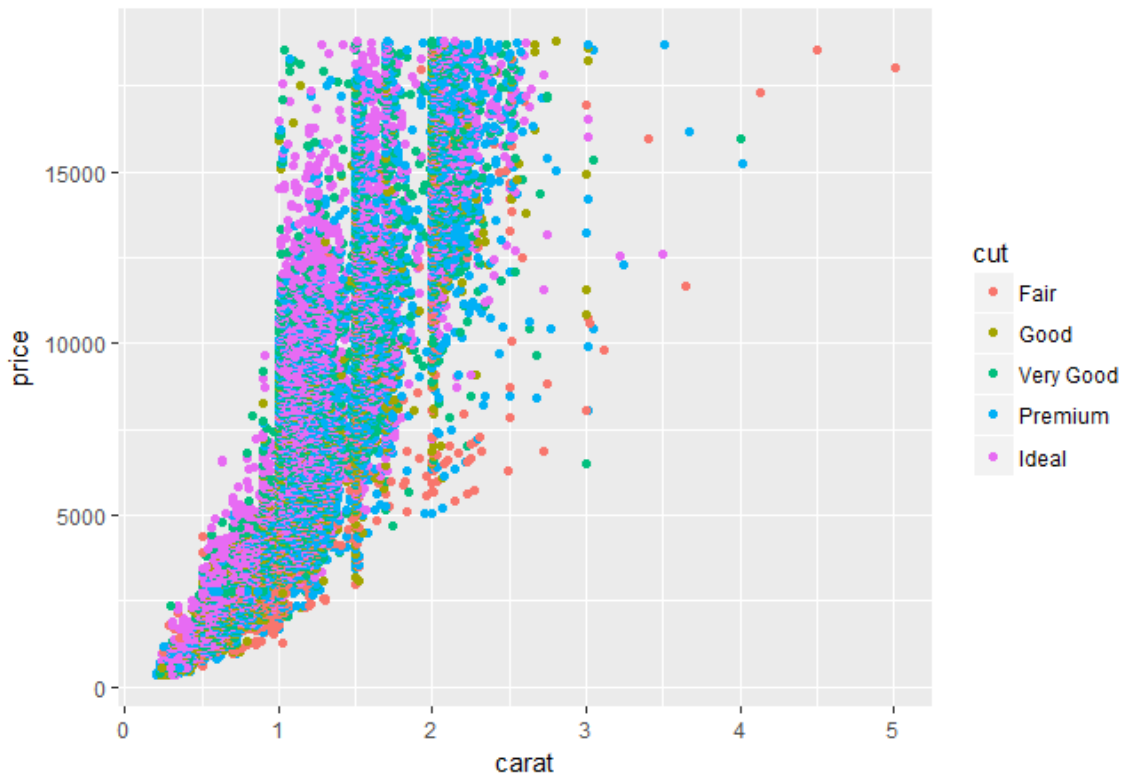
```
ggplot(data=diamonds, aes(x=carat, y=price)) +
  geom_point(col="lightblue")
```



Знайдіть коефіцієнт кореляції між вагою (змінна **carat**) та ціною (змінна **price**) діамантів. Результат вкажіть в якості відповіді на питання 3.

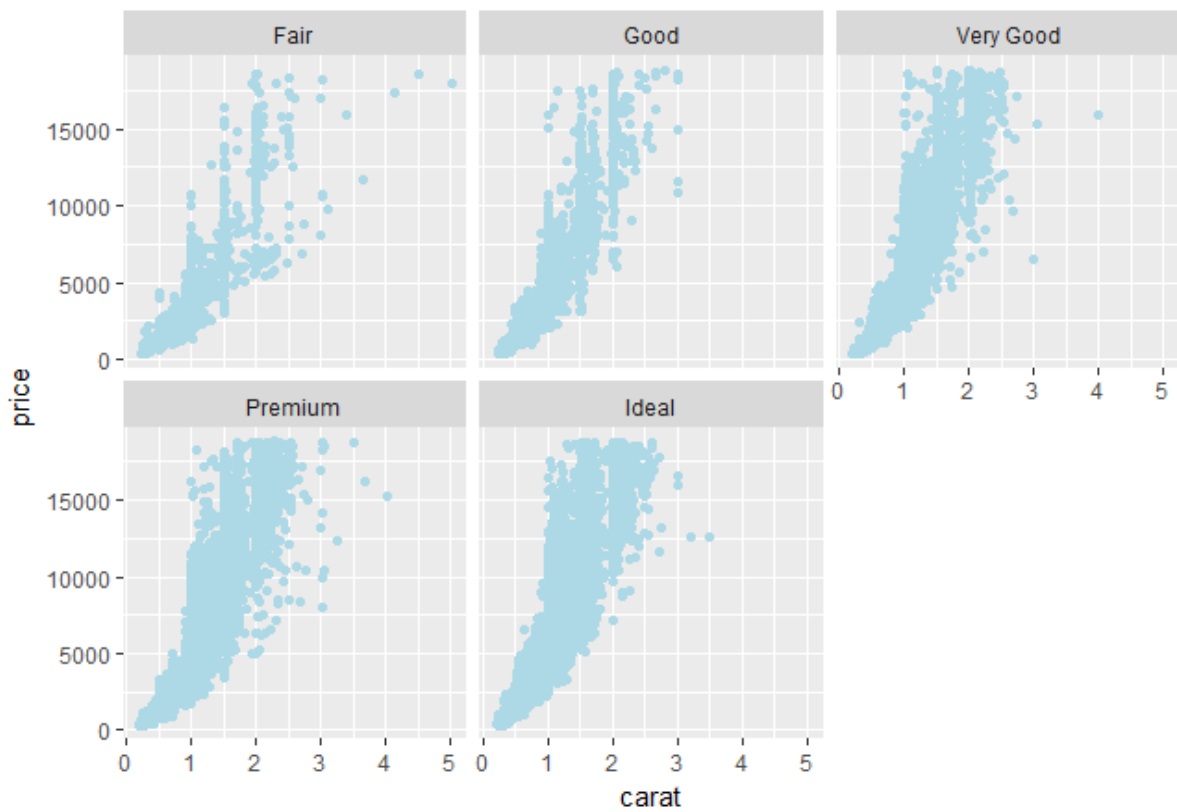
Давайте подивимось, як розподілені вага та ціна в залежності від ступеня обробки діамантів:

```
ggplot(data=diamonds, aes(x=carat, y=price, col=cut)) +  
  geom_point()
```



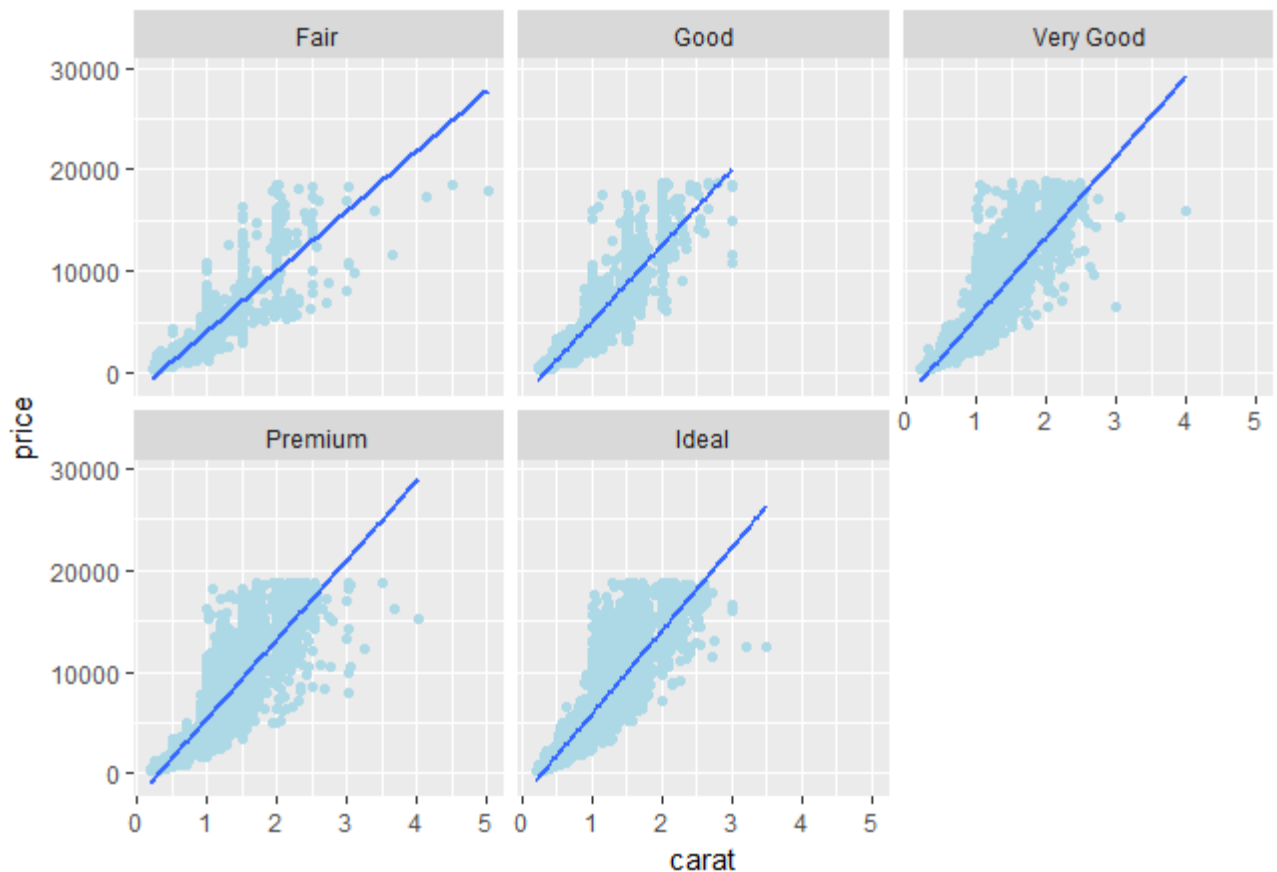
Інший тип відображення:

```
ggplot(data=diamonds, aes(x=carat, y=price)) +  
  geom_point(col="lightblue") +  
  facet_wrap(~cut)
```



Додамо до графіка ще лінію лінійної регресії:

```
ggplot(data=diamonds, aes(x=carat, y=price)) +  
  geom_point(col="lightblue") +  
  geom_smooth(method="lm", se=FALSE) +  
  facet_wrap(~cut)
```



Побудуйте моделі лінійної регресії `lin.diamond.ideal` та `lin.diamond.fair` залежності ціни від ваги для обробки (змінна `cut`) `Ideal` та `Fair` відповідно. Знайдіть ціну ідеально та прийнятно обробленого діаманта вагою 1 карат згідно побудованих лінійних моделей. Вкажіть знайдені значення в якості відповіді на питання 4 та 5 відповідно.

Завдання:

Опрацювати теоретичний матеріал та **виконати** всі **описані** завдання лабораторної роботи (зберегти R Скрипт).

Дати відповідь на наступні запитання:

1. Який коефіцієнт кореляції для квартету Анскомбе?

- a) 0.453 b) 0.816 c) 0.716 d) -0.743

2. Яке рівняння лінійної регресії для квартету Анскомбе?

- a) $\hat{y} = -0.5x + 3$ b) $\hat{y} = -0.5x - 3$ c) $\hat{y} = 0.5x + 3$ d) $\hat{y} = 3x + 0.5$

3. Чому дорівнює коефіцієнт кореляції для змінних ціна та вага з набору diamonds?

- a) -0.64 b) 0.75 c) 0.35 d) 0.92

4. Якою буде ціна ідеально (cut = Ideal) обробленого діаманта вагою 1 карат згідно моделі lin.diamond.ideal?

- a) 5432 b) 7134 c) 2145 d) 5892

5. Якою буде ціна прийнятно (cut = Fair) обробленого діаманта вагою 1 карат згідно моделі lin.diamond.fair?

- a) 5892 b) 6000 c) 3567 d) 4085

Самостійно провести діагностику лінійних моделей регресії lin.diamond.ideal та lin.diamond.fair (зберегти R Скрипт).