# *Оцінка параметрів рівняння лінійної регресії*

**Мета роботи**: ознайомитись з методикою оцінки параметрів рівняння прямої лінії середньоквадратичної регресії; випробувати її для прогнозування.

***Основне завдання***

1. Визначити точкові оцінки параметрів рівняння лінійної регресії, записати отримане рівняння та побудувати графік.
2. Зробити прогноз для кількох значень за побудованою моделлю та порівняти його з фактичними значеннями.

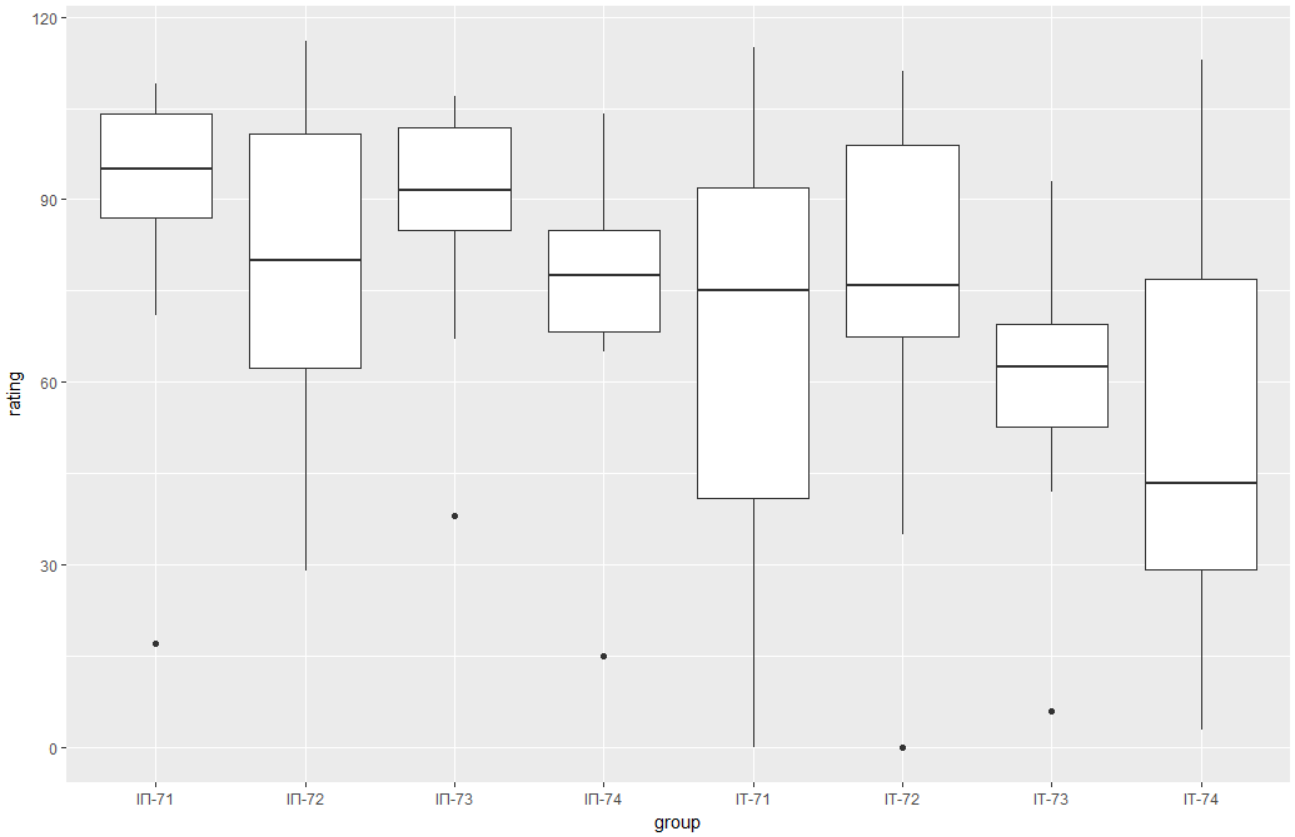
**Код програми та отримані результати:**

Дослідимо дані про успішність студентів, що навчаються в різних групах. Спробуємо визначити, що впливає на загальний рейтинг студента. Які впливи є найсуттєвішими, а які мають незначний вплив на успішність?

```
library(readr)
data = read.csv("laba7.csv", sep = ";")
data. = data [,c(1,3:14)]       #without sernames
data.numeric = data[,c(9:14)]   #only numeric columns
View(data.)
```

| | num | group | theme1 | theme2 | theme3 | theme4 | theme5 | ask | dkr | control | themes | absense | rating |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 19 | 19 | IT-71 | 3 | 6 | 3 | 6 | 0 | 3 | 14 | 60 | 18 | 7 | 98 |
| 20 | 20 | IT-71 | 3 | 0 | 6 | 9 | 0 | 0 | 14 | 39 | 18 | 9 | 71 |
| 21 | 21 | IT-71 | 0 | 0 | 0 | 0 | 0 | 3 | 8 | 7 | 0 | 16 | 18 |
| 22 | 22 | IT-71 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 17 | 0 | 18 | 17 |
| 23 | 23 | IT-71 | 0 | 3 | 3 | 3 | 0 | 3 | 14 | 49 | 9 | 11 | 75 |
| 24 | 24 | IT-71 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 21 | 0 |
| 25 | 25 | IT-71 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 17 | 6 |
| 26 | 26 | IT-71 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 8 | 0 | 19 | 11 |
| 27 | 27 | IT-71 | 0 | 3 | 0 | 6 | 3 | 3 | 15 | 66 | 12 | 10 | 96 |
| 28 | 28 | IT-71 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 13 | 0 | 19 | 13 |
| 29 | 29 | IT-71 | 3 | 3 | 6 | 9 | 3 | 3 | 15 | 56 | 24 | 5 | 101 |
| 30 | 30 | IT-72 | 3 | 6 | 3 | 0 | 0 | 3 | 13 | 52 | 12 | 10 | 80 |
| 31 | 31 | IT-72 | 0 | 0 | 0 | 3 | 3 | 3 | 15 | 43 | 6 | 11 | 70 |
| 32 | 32 | IT-72 | 3 | 3 | 6 | 9 | 1 | 3 | 15 | 35 | 22 | 6 | 75 |
| 33 | 33 | IT-72 | 0 | 3 | 3 | 3 | 0 | 3 | 14 | 50 | 9 | 11 | 76 |
| 34 | 34 | IT-72 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 21 | 0 |
| 35 | 35 | IT-72 | 0 | 0 | 0 | 0 | 0 | 0 | 15 | 39 | 0 | 15 | 54 |
| 36 | 36 | IT-72 | 0 | 3 | 6 | 0 | 0 | 3 | 14 | 43 | 9 | 10 | 72 |
| 37 | 37 | IT-72 | 3 | 0 | 0 | 0 | 1 | 3 | 14 | 44 | 4 | 12 | 65 |
| 38 | 38 | IT-72 | 3 | 0 | 3 | 6 | 0 | 3 | 14 | 58 | 12 | 9 | 90 |
| 39 | 39 | IT-72 | 0 | 3 | 3 | 0 | 0 | 3 | 14 | 18 | 6 | 14 | 41 |
| 40 | 40 | IT-72 | 3 | 6 | 9 | 9 | 0 | 3 | 15 | 63 | 27 | 4 | 111 |
| 41 | 41 | IT-72 | 0 | 3 | 9 | 6 | 0 | 3 | 15 | 59 | 18 | 7 | 98 |

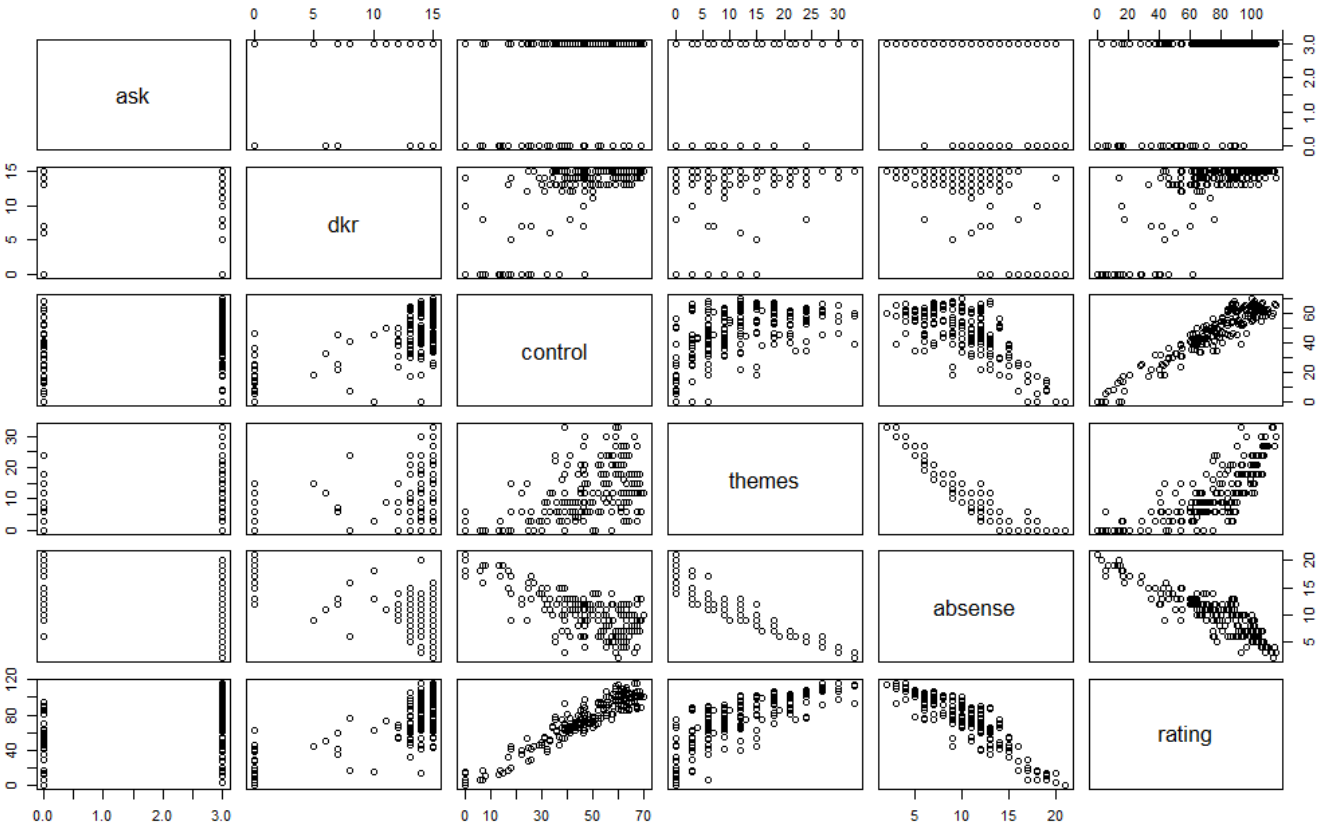
```
ggplot(data, aes(x = group, y = rating)) +
  geom_boxplot()
```




Можна помітити, що успішність студентів ІП дещо вище, ніж ІТ.

Проведемо попарний аналіз числових величин, щоб визначити, які фактори впливають на успішність студентів.

```
pairs(data.numeric) ## попарный анализ
```



Можна побачити, що відносно сильно лінійно впливають на успішність (rating) лише деякі фактори: пропуски занять(absense), набрані на заняттях бали (themes) та контрольні роботи (control). Причому, видно, що контрольні та бали на заняттях впливають впливають на успішність з прямою залежністю, а пропуски – з оберненою. Для характеристики даної залежності використовують коефіцієнт кореляції:

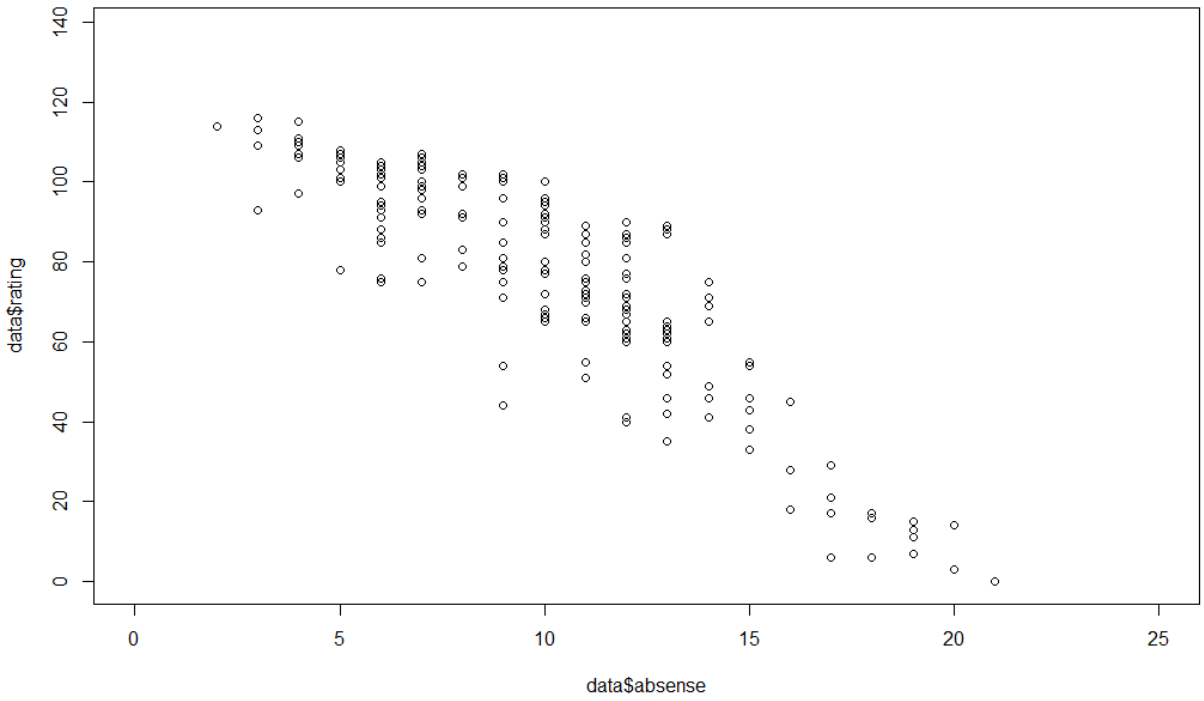
```
> cor.test(x = data$rating, y = data$absense) ## отрицательная зависимость

        Pearson's product-moment correlation

data:  data$rating and data$absense
t = -28.031, df = 207, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9149051 -0.8574712
sample estimates:
       cor
-0.8896554

> cor.test(x = data$rating, y = data$themes)  ## положительная зависимость

        Pearson's product-moment correlation

data:  data$rating and data$themes
t = 16.939, df = 207, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6987429 0.8137241
sample estimates:
      cor
0.7621803
```

Знайдемо коефіцієнти кореляції та p-значення при попарному порівнянні величин:

```
> corr = corr.test(data.numeric)
> corr$r      ## построение корреляционной матрицы
               ask        dkr    control     themes    absense     rating
ask      1.0000000  0.4619574  0.4204317  0.3272097 -0.4858318  0.4967374
dkr      0.4619574  1.0000000  0.6809528  0.4157386 -0.5985987  0.7525810
control  0.4204317  0.6809528  1.0000000  0.5353191 -0.7073655  0.9419795
themes   0.3272097  0.4157386  0.5353191  1.0000000 -0.9408554  0.7621803
absense -0.4858318 -0.5985987 -0.7073655 -0.9408554  1.0000000 -0.8896554
rating   0.4967374  0.7525810  0.9419795  0.7621803 -0.8896554  1.0000000
> corr$p
                ask          dkr       control       themes      absense       rating
ask     0.000000e+00 7.652581e-12  6.990351e-10 1.325920e-06 4.443022e-13 1.209609e-13
dkr     1.913145e-12 0.000000e+00  7.180726e-29 7.705462e-10 8.378734e-21 2.226916e-38
control 2.330117e-10 7.978585e-30  0.000000e+00 4.753065e-16 5.037614e-32 6.445200e-99
themes  1.325920e-06 3.852731e-10  6.790093e-17 0.000000e+00 4.133138e-98 7.029651e-40
absense 8.886045e-14 1.047342e-21  5.037614e-33 2.952241e-99 0.000000e+00 2.741166e-71
rating  2.016015e-14 2.024469e-39 4.296800e-100 5.858043e-41 2.108590e-72 0.000000e+00
>
```

Чим менше p-значення та чим ближче коефіцієнт кореляції між величинами, тим сильніший між ними зв’язок (тим більший вплив здійснює одна величина на іншу).

Далі проведемо дослідження впливу пропусків на успішність студентів:




```
lm.data = lm(rating ~ absense, data) ## построение линейной регрессии
a = summary(lm.data) ## rat = -6*abs +137.5
```

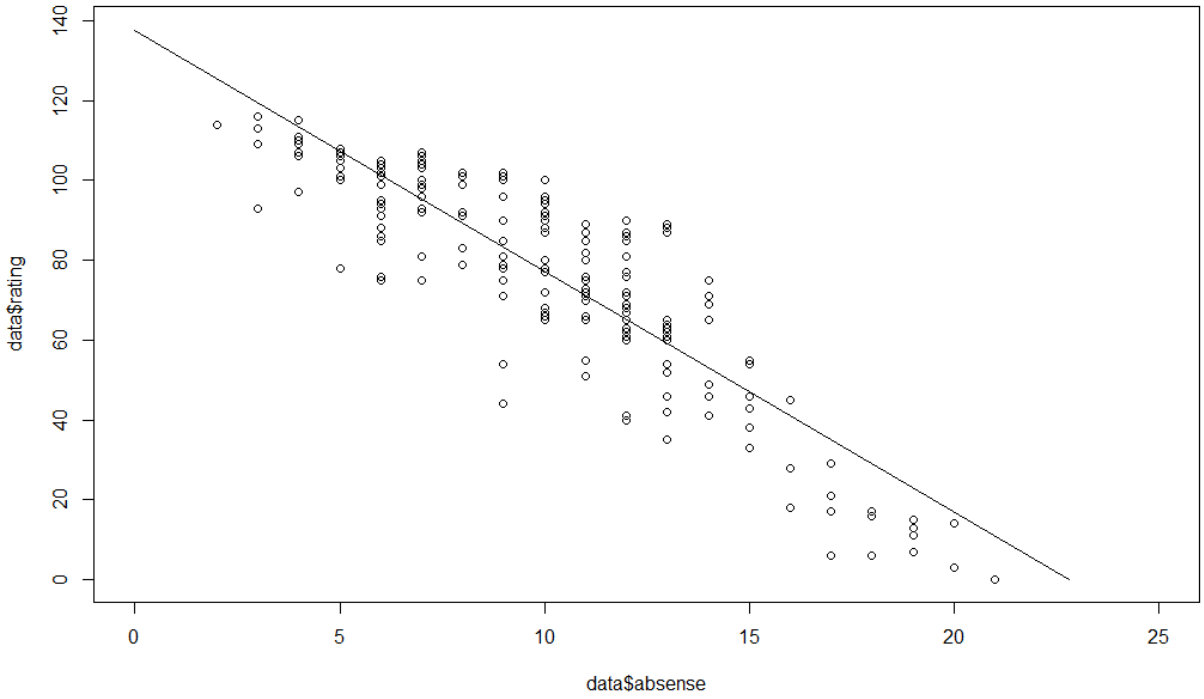
```
> a = summary(lm.data) ## rat = -6*abs +137.5
> a

Call:
lm(formula = rating ~ absense, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-39.249  -7.340   0.629   8.690  29.873

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 137.5237     2.3334   58.94   <2e-16 ***
absense      -6.0305     0.2151  -28.03   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.33 on 207 degrees of freedom
Multiple R-squared:  0.7915,    Adjusted R-squared:  0.7905
F-statistic: 785.7 on 1 and 207 DF,  p-value: < 2.2e-16
```

Чим ближче значення Adjusted R-squared до одиниці, тим краще дана модель буде описувати залежність.

Визначаємо точкові оцінки параметрів рівняння лінійної регресії, записуємо отримане рівняння : ***rating = -6*absanse +137.5***

Побудуємо графік за двома точками:

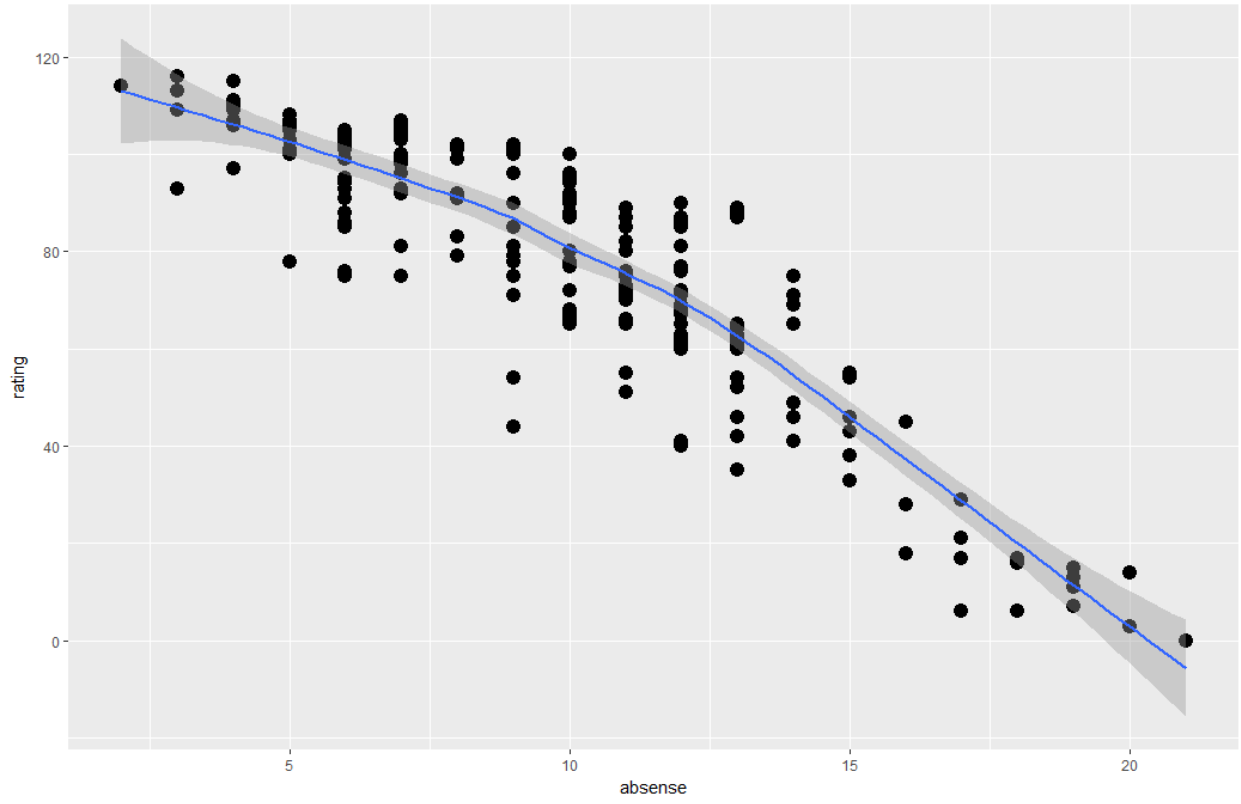
```
free = a$coefficients[1];coef = a$coefficients[2]

abs0 = 0; rat = coef*abs0 + free

rat0 = 0; abs = (rat0 - free)/coef
lines(c(abs0,abs),c(rat,rat0))
```




Бачимо, що лінія достатньо непогано описує дану діаграму розсіювання.

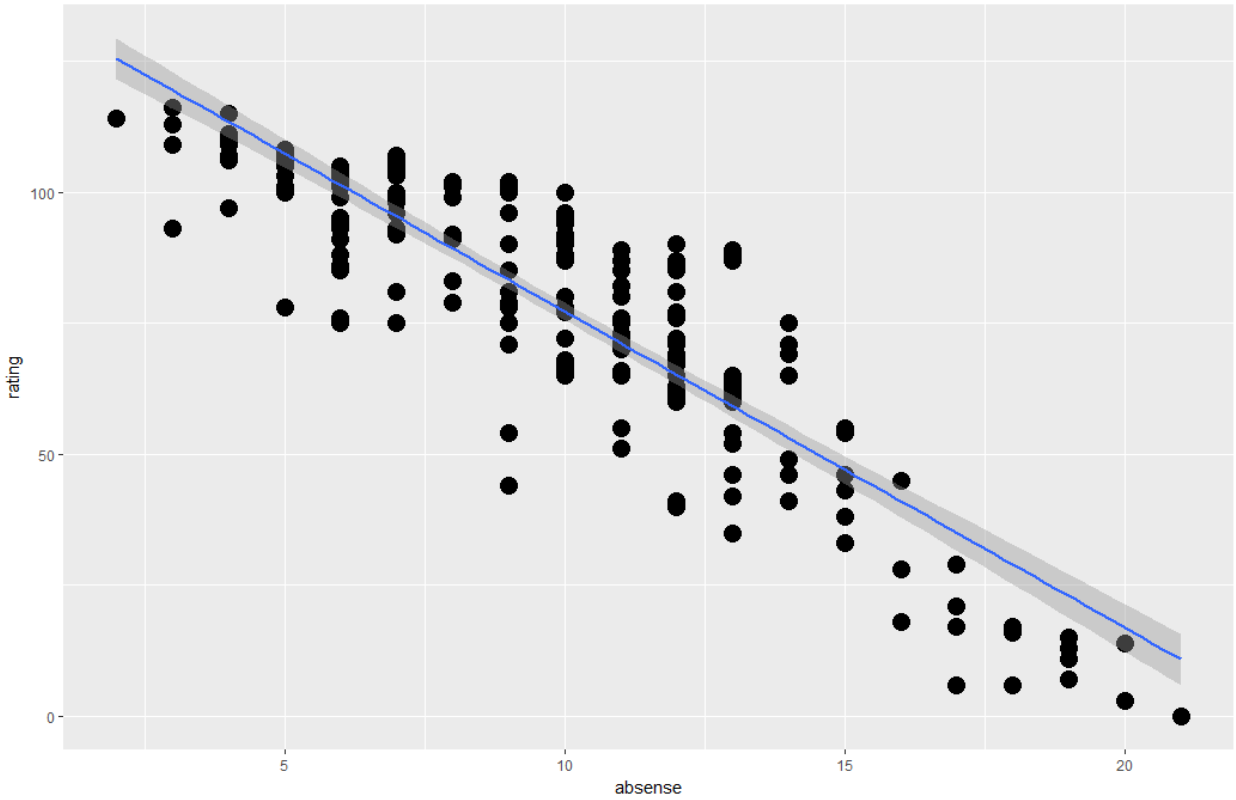
Знайдемо дану лінію тренду програмними засобами бібліотеки ggplot2:

```
#smoothing
ggplot(data, aes(absense, rating))+
  geom_point(size = 4)+
  geom_smooth()
```



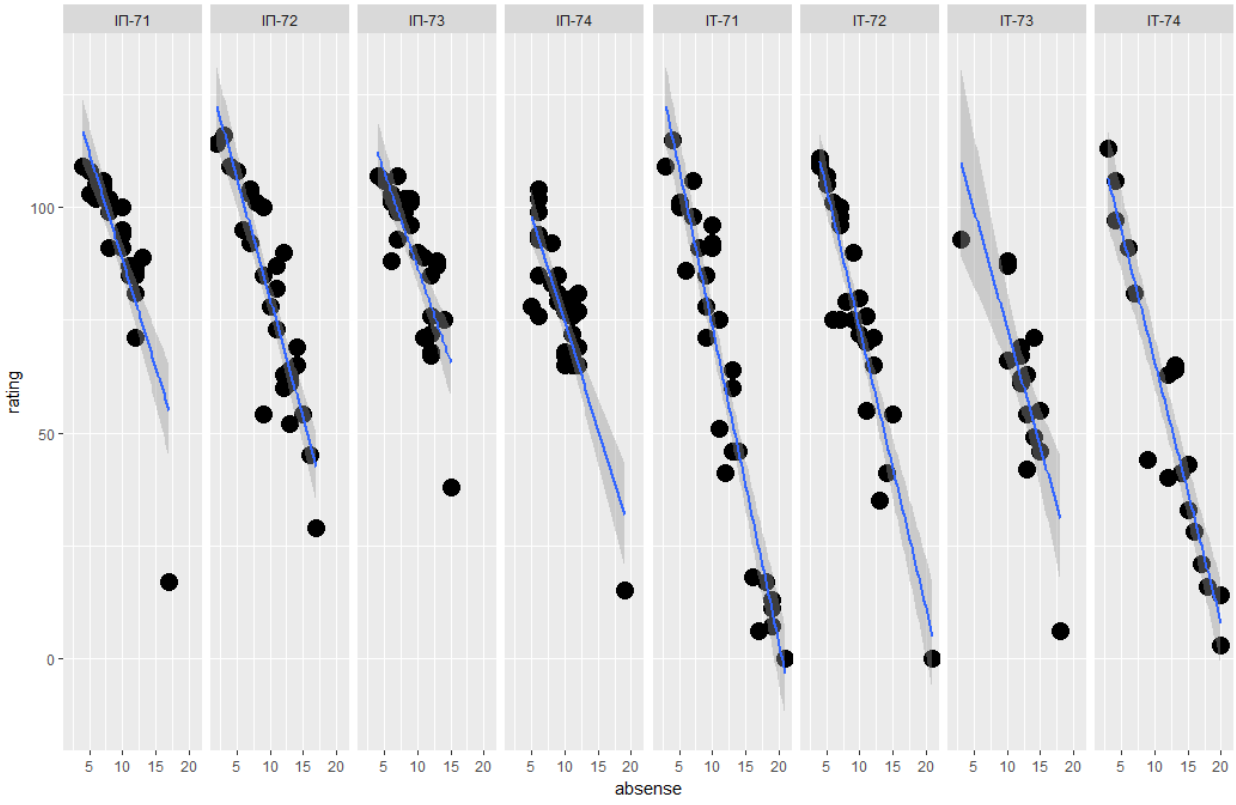

```
ggplot(data, aes(absense, rating))+
  geom_point(size = 5)+
```

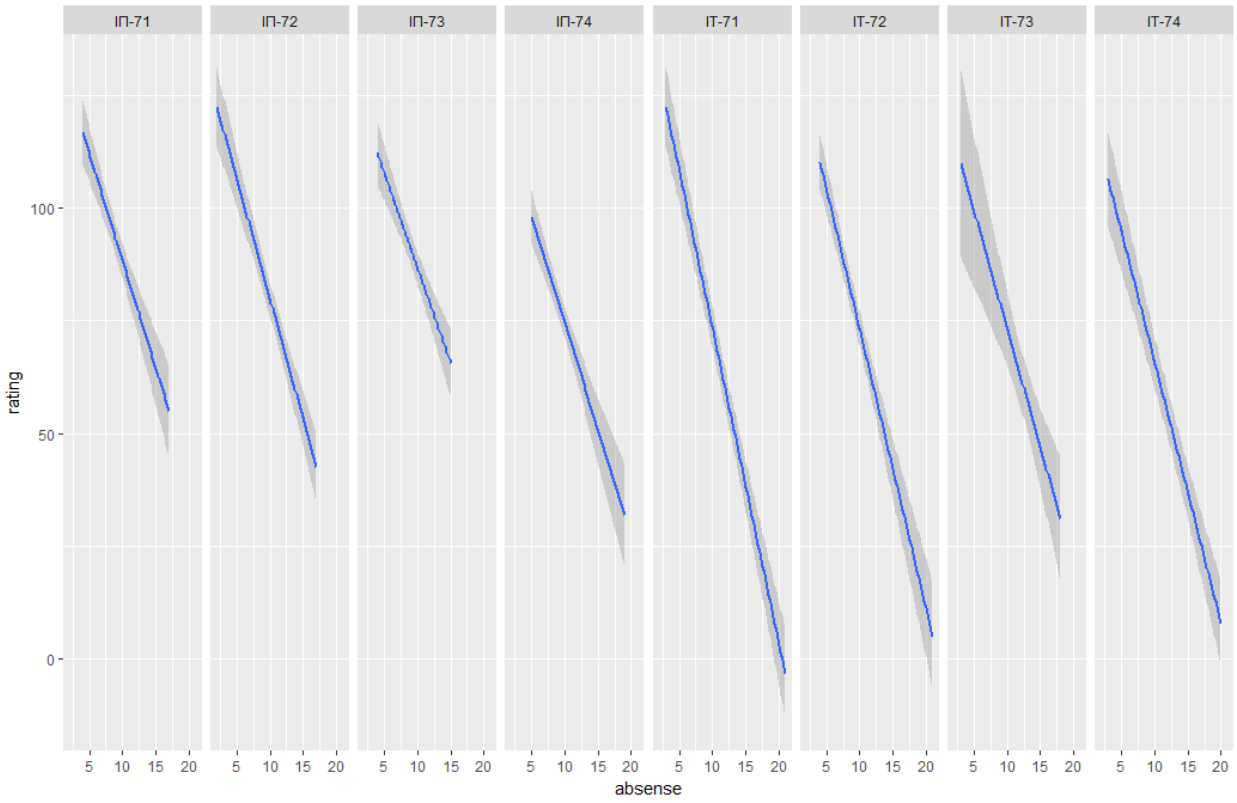
```
geom_smooth(method = "lm") #smoothing line
```




Можна бачити, що лінія співпадає з тією, що задана рівнянням. Отже, знайдене рівняння – правильне.

```
ggplot(data, aes(absense, rating))+
  geom_point(size = 5)+
  geom_smooth(method = "lm")+
  facet_grid(.~group) #smoothing line on groups
```






Можна бачити, що лінія тренду мало відрізняється у різних груп.

Створимо таблицю для порівняння реальних даних успішності з прогнозованими за раніше проведеним дослідженням допомогою аргументу fitted.values.

```
fitted_values_rating = data.frame( student = data$num, rating = data$rating, fitted = lm.data$fitted.values)
```

Бачимо, що деякі спрогнозовані значення збігаються с реальними, але є й такі що суттєво відрізняються. Це може буде спричинено тим, що існують сторонні фактори, що впливають на успішність.

fitted_values_rating

Filter

| | student | rating | fitted |
|---|---|---|---|
| 1 | 1 | 91 | 77.21823 |
| 2 | 2 | 115 | 113.40149 |
| 3 | 3 | 51 | 71.18769 |
| 4 | 4 | 100 | 107.37095 |
| 5 | 5 | 92 | 77.21823 |
| 6 | 6 | 106 | 95.30986 |
| 7 | 7 | 46 | 53.09606 |
| 8 | 8 | 86 | 101.34041 |
| 9 | 9 | 91 | 89.27932 |
| 10 | 10 | 85 | 83.24878 |
| 11 | 11 | 64 | 59.12660 |
| 12 | 12 | 41 | 65.15715 |
| 13 | 13 | 7 | 22.94335 |
| 14 | 14 | 46 | 59.12660 |
| 15 | 15 | 60 | 59.12660 |
| 16 | 16 | 109 | 119.43204 |
| 17 | 17 | 78 | 83.24878 |
| 18 | 18 | 91 | 77.21823 |
| 19 | 19 | 98 | 95.30986 |
| 20 | 20 | 71 | 83.24878 |
| 21 | 21 | 18 | 41.03498 |
| 22 | 22 | 17 | 28.97389 |
| 23 | 23 | 75 | 71.18769 |
| 24 | 24 | 0 | 10.88226 |
| 25 | 25 | 6 | 35.00443 |
| 26 | 26 | 11 | 22.94335 |
| 27 | 27 | 96 | 77.21823 |

Спробуємо спрогнозувати успішність студента за відомою кількістю занять, що він пропустив. Прогнозування здійснюється за допомогою функції predict().

```
new_absense = data.frame(absense = c(2,5,10,20))
new_absense$rating = predict(lm.data, new_absense)
```

| | absense | rating |
|---|---|---|
| 1 | 2 | 125.46258 |
| 2 | 5 | 107.37095 |
| 3 | 10 | 77.21823 |
| 4 | 20 | 16.91280 |

**Висновок**

Дане дослідження показало, що на успішність студента явним чином впливають наступні чинники: пропуски занять, набрані на заняттях бали та контрольні роботи. Також, судячи с отриманих коефіцієнтів кореляції, можемо прийти висновку, що пропуски занять негативно впливають на рейтинг на відміну від балів за контрольні роботи та ті, що були набрані на заняттях.

Отримавши рівняння лінійної регресії залежності рівня успішності від кількості пропусків, я отримала змогу прогнозувати загальний бал студента за заздалегідь відомим числом пропущених занять. Порівнявши, спрогнозовані дані з реальними, я вважаю, що отримана мною модель є коректною та досить точною.

Також в ході дослідження я помітила цікавий факт, якого досі не знала: студенти з потоку ІП мають більш високу успішність ніж з потоку ІТ.