Лабораторна робота №2
з дисципліни
”регресійний аналіз та асимптотична статистика”
Варіант №4

Горбунова Даніела Денисовича
4 курс бакалаврату
група "комп’ютерна статистика"

6 березня 2021 р.

1 Вступ.

У даній роботі вказана інформація про отримані результати під час виконання роботи №2: сформовано дані для статистичного дослідження, побудовано регресійні моделі, знайдено оцінки коефіцієнтів регресії та досліджені залишки. Визначена значущість регресорів, побудовані довірчі інтервали та знайдені дисперсії похибок вищезгаданих оцінок.

1.1 Початкові дані.

Маємо простий набір даних: результати модульних контрольних робіт студентів 3 курсу з математичної статистики. Питання таке: чи наявна залежність між відповідними результатами модульних контрольних робіт з математичної статистики (див. додаток 4.1); чи можна на основі визначеної залежності спрогнозувати нові результати? Для цього побудуємо регресійну модель.

1.2 Дескриптивна статистика даних.

Побудуємо діаграму розсіювання даних. Маємо таку картину: З діаграми видно, що між

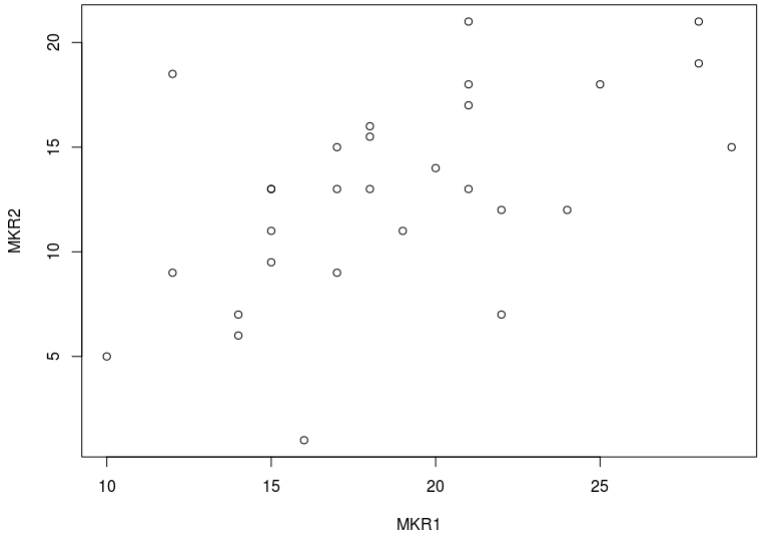


Рис. 1: Діаграма розсіювання даних.

результатами контрольних робіт виражається залежність, схожа на лінійну. Значущими є і коефіцієнти кореляції Пірсона та Спірмена:

$$corr_{pearson}(MKR1, MKR2) = 0.5474906$$
$$corr_{spearman}(MKR1, MKR2) = 0.5277262$$

Додатково наводяться гістограми відносних частот $MKR1$ та $MKR2$:

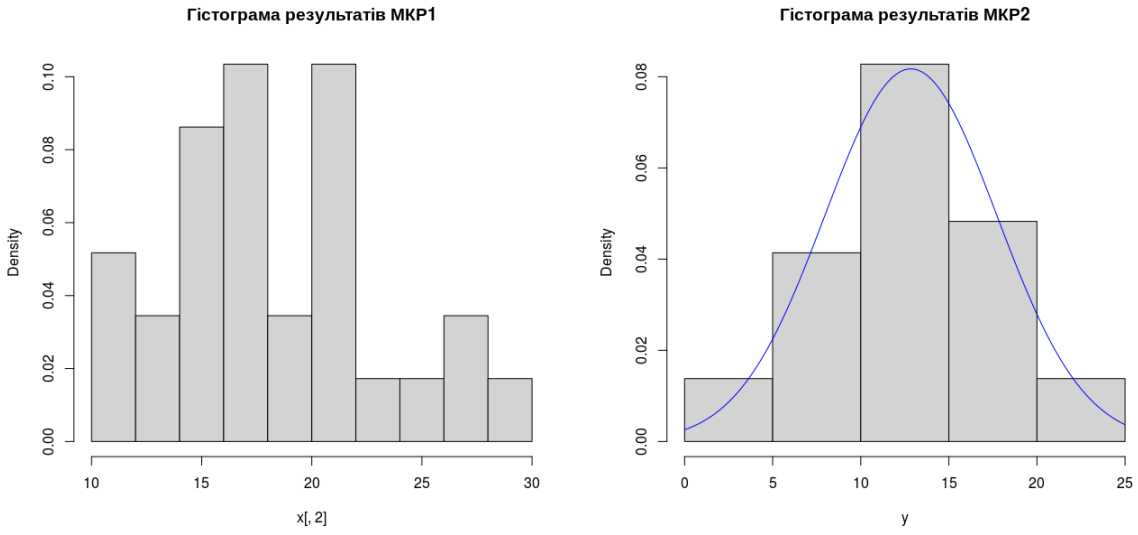


Рис. 2: Гістограми результатів контрольних робіт.

2 Хід роботи.

2.1 Побудова регерсійної моделі.

В якості регресора X та відгука Y беремо результати першої та другої МКР відповідно. Маємо лінійну модель:

$$Y_j = b_1 + b_2 X_j, j = \overline{1, N}$$

де вектор $b = (b_1, b_2)^\top$ знайдемо з нормального рівняння регресії:

$$Ab = X^\top Y,\ A = X^\top X \tag{1}$$

Зауважимо, що

$$A = X^\top X = \begin{pmatrix} 29 & 544 \\ 544 & 10878 \end{pmatrix},\ \det A = 19526 \neq 0$$

тому рівняння (1) має єдиний розв'язок:

$$\hat{b} = A^{-1} X^\top Y = \begin{pmatrix} 0.55710335 & -0.027860289 \\ -0.02786029 & 0.001485199 \end{pmatrix} \begin{pmatrix} 372.5 \\ 7354.5 \end{pmatrix} = \begin{pmatrix} 2.6225033 \\ 0.5449401 \end{pmatrix}$$

який буде оцінкою за методом найменших квадратів. Оцінку параметрів моделі знайшли, тепер варто дослідити її.

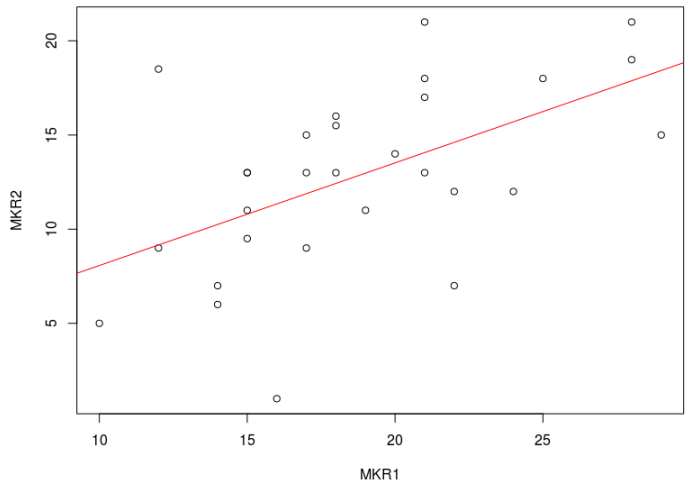


Рис. 3: Діаграма розсіювання та пряма регресії.

2.2 Дослідження оцінки МНК.

Надалі залишки в регресійній моделі позначимо через $U = Y - \hat{Y}$, де $Y = X\hat{b}$.

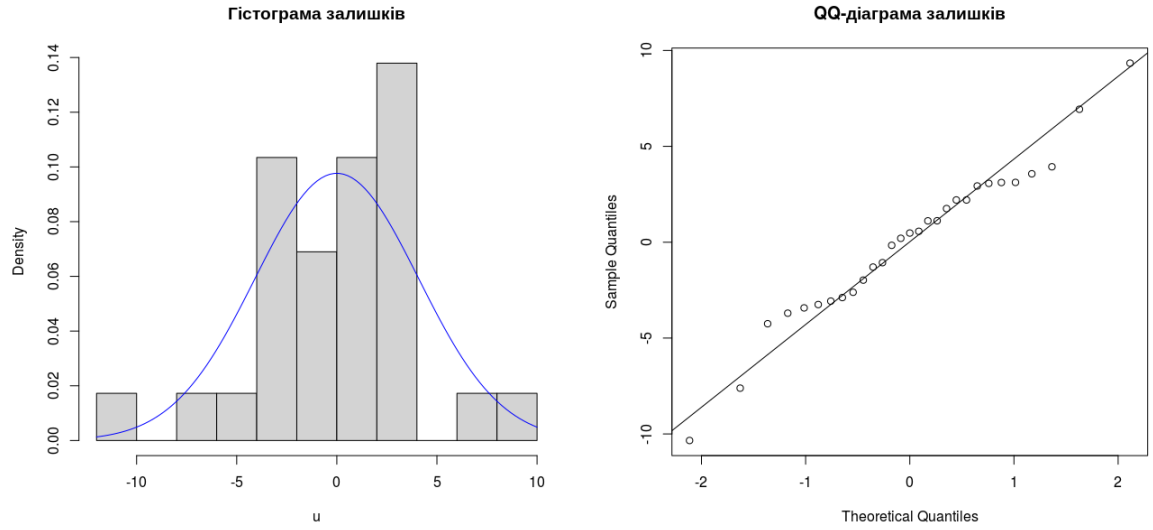
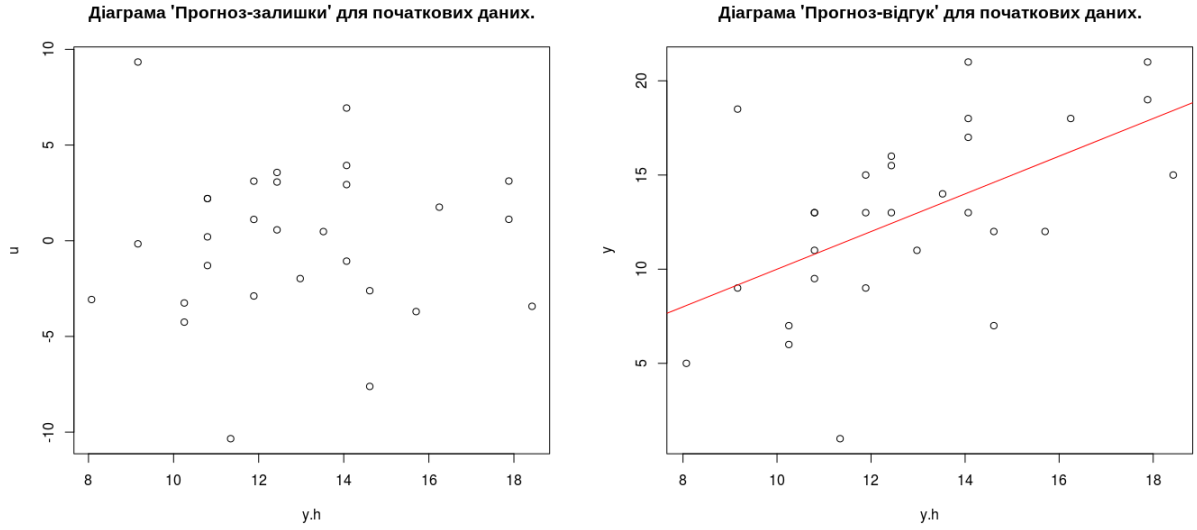


Рис. 4: Гістограма та QQ-діаграма залішків регресійної моделі. На гістограмі підігнана крива щільності гауссового розподілу.

Оскільки кількість спостережень, що використовуються для побудови моделі, незначна, тому точних висновків про гауссовість розподілу позибок говорити не можна. Видно, що на гістограмі відносних частот твориться справжній хаос, а на квантильній діаграмі помітні відхилення є на кінцях. Розкид похибок є досить високим.



Діаграма "прозноз-залишки" натякає на те, що можна розглянути іншу модель без використання деяких точок.

Діаграма "прогноз-відгук" показує, що прогноз вловлює лінійну тенденцію, хоч і з великим розкидом, як було зазначено раніше.

Обчислимо незміщену оцінку дисперсії похибок моделі:

$$\hat{\sigma}^2 = \frac{1}{N-d}\sum_{j=1}^{N} U_j^2 = \frac{1}{N-2}\sum_{j=1}^{N} U_j^2 = 17.30021$$

За допомогою вищенаведеної оцінки, можна оцінити дисперсії оцінок коефіцієнтів регресії:

$$\hat{\sigma}^2_{\hat{b}_i} = \hat{\sigma}^2 (A^{-1})_{ii} = \begin{cases} 9.63800540, & i = 1 \\ 0.02569426, & i = 2 \end{cases}$$

Тепер можна перевірити певні гіпотези та побудувати довірчі інтервали для коефіцієнтів моделі (надалі рівень значущості $\alpha = 0.05$).

Чи має місце залежність відгуку від регресора? Для цього використаємо тест Фішера для перевірки загальної лінійної гіпотези, сформулювавши такі статистичні гіпотези:

$$H_0 : b_2 = 0$$
$$H_1 : b_2 \neq 0$$

У такому разі, статистика тесту набуває такого вигляду ($m = d - 1$):

$$F_{emp} = \frac{N-m-1}{m}\frac{R^2}{1-R^2} = 63.07629, \text{ де}$$
$$R^2 = 1 - \frac{\sum_{j=1}^{N} U_j^2}{\sum_{j=1}^{N}(Y_j - \overline{Y})^2} = 0.2997459$$

А квантиль розподілу Фішера з ступенями вільності $(m, N-m-1)$ дорівнює:

$$F_{theor} = Q^{F_{m,N-m-1}}(1-\alpha) = 4.210008 < F_{emp}$$

Отже при $\alpha = 0.05$ слід прийняти альтернативу (залежність від регресора наявна).

Чи є суттєвою відмінність коефіцієнтів регресії від нуля? Вводимо та обчислюємо t-статистику:

$$\hat{t}_i = \frac{\hat{b}_i}{\hat{\sigma}_{\hat{b}_i}} = \begin{cases} 0.8447388, & i = 1 \\ 3.3996225, & i = 2 \end{cases}$$

Відповідні ймовірності на хвостах дорівнюють:

$$\mathbb{P}\left(|t_{N-d}| \geqslant \hat{t}_i\right) = 1 - \mathbb{P}\left(|t_{N-d}| < \hat{t}_i\right) = 2\mathbb{P}\left(t_{N-d} < \hat{t}_i\right) = \begin{cases} 0.405679990, & i = 1 \\ 0.002112383, & i = 2 \end{cases}$$

При заданому α слід вивести, що коефіцієнт нахилу в побудованій моделі, тобто b_2, значущо відрізняється від нуля.

Які довірчі інтервали для коефіцієнтів регресії? За отриманими оцінками МНК, використаємо попередні результати для побудови інтервалів рівня $1 - \alpha = 0.95$:

$$\text{Для } b_i : [\hat{b}_i - \hat{\sigma}_{\hat{b}_i} Q^{t_{N-d}}(1-\alpha/2), \hat{b}_i + \hat{\sigma}_{\hat{b}_i} Q^{t_{N-d}}(1-\alpha/2)],$$
$$\text{де } Q^{t_{N-d}}(1-\alpha/2),\ Q^{t_{N-d}}(1-\alpha/2) := 2.051831$$
$$\text{Конкретно: для } b_1 : [-3.7474326,\ 8.9924393];\ \text{для } b_2 : [0.2160434,\ 0.8738368]$$

3 Висновки.

На результатах контрльних робіт з математичної статистики можна використати лінійну регресію. Показники моделі вийшли не зовсім хорошими, але й не можна вважати поганими. Адекватність прогнозу є сумнівною: однією з можливих причин є невеликий обсяг даних.

Слушною думкою було б погратися з вилученням спостережень, які могли б слугувати в якості викидів, але в даній роботі поки що це не наводиться.

4 Додаток.

4.1 Використаний набір даних.

Дані за 2020 рік, вказані лише результати без переписування контрольних робіт. За першу роботу максимально можна було отримати 30 балів, за другу - 21. Усього $N = 29$ спостережень.

МКР1	МКР2
15	13
12	18.5
22	12
21	18
28	19
18	13
15	11
21	13
21	21
25	18
24	12
22	7
17	15
18	16
17	9
14	7
28	21
14	6
21	17
16	1
19	11
15	9.5
29	15
17	13
15	13
20	14
18	15.5
12	9
10	5

Рис. 5: Результати модульних контрольних робіт з математичної статистики.

4.2 Програмна реалізація.

```
# Реалізація та аналіз першої моделі. Для другої аналогічно
df <- read.csv("ms3.csv", sep = ' ')[c('MKR1', 'MKR2')]
plot(df)
df <- df[-10,]
row.names(df) <- NULL
n <- nrow(df)
d <- 2

x <- cbind(1 + numeric(n), data.matrix(df[1]))
y <- data.matrix(df[2])

hist(x[,2], probability = T, main = "Гістограма результатів МКР1", breaks = 10)
hist(y, probability = T, main = "Гістограма результатів МКР2")
curve(dnorm(x, mean(y), sd(y)), col = 'blue', add = T)

a <- t(x)%*%x
a.inv <- solve(a)
b <- a.inv%*%t(x)%*%y # LS estim

plot(df)
abline(b, col = 'red')

y.h <- x%*%b # predictions
u <- y - y.h # residuals

# descriptive statistics
hist(u, probability = T, breaks = 10, main = "Гістограма залишків")
curve(dnorm(x, mean(u), sd(u)), col = 'blue', add = T)

qqnorm(u, main = "QQ-діаграма залишків")
qqline(u)

plot(y.h, u, main = "Діаграма 'Прогноз-залишки' для початкових даних.")
plot(y.h, y, main = "Діаграма 'Прогноз-відгук' для початкових даних.")

# r-squared
r.sq <- sum(u^2)/sum((y - mean(y))^2)
print(1 - r.sq)

# variance of errors estimation
err.var.estim <- sum(u^2)/(n - d)
print(err.estim)

# variance of coefficients estimation
coef.var.estim <- err.var.estim * diag(a.inv)
print(sqrt(coef.var.estim))
```

```
# t-test
t.coef.0 <- -b/sqrt(coef.var.estim)
print(t.coef.0)
alpha <- 0.05
q.t <- qt(1 - alpha/2, df = n - d)
print(2*pt(t.coef.0, df = n - d))

err.t <- q.t * sqrt(coef.var.estim)

# c.i. for b
l.b <- b - err.t
u.b <- b + err.t
print("c.i.")
print(l.b)
print(u.b)
```