Міністерство освіти і науки України
НТУУ «КПІ ім. Ігоря Сікорського»
Навчально-науковий інститут атомної та теплової енергетики
Кафедра цифрових технологій в енергетиці

Лабораторна робота №1

з дисципліни «Комп’ютерне моделювання»

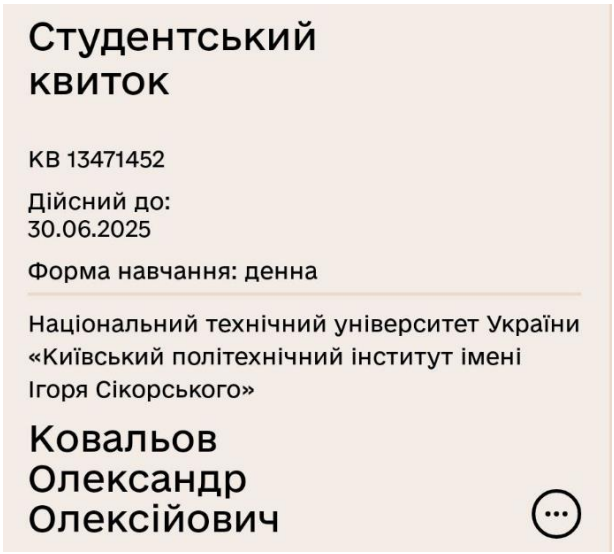
Тема «Статистичне моделювання»

Варіант №22

Студента 3-го курсу НН ІАТЕ гр. ТР-12
Ковальова Олександра
Перевірив: д.т.н., проф. Шушура О. М.

КИЇВ 2023

Варіант. Для наведеної задачі та таблиці даних по своєму варіанту (номер варіанту дорівнює g, де g – остання цифра у номері студентського квитка + 1).



Остання цифра у номері студентського квитка – 2, варіант роботи – **3**.

Загальне завдання.

1. Перевірити дані на наявність помилок та виключити їх (обґрунтувати застосування методів виключення);
2. Побудувати гістограму по кожному стовпцю даних в таблиці;
3. Оцінити взаємний вплив характеристик, наведених в таблиці;
4. За допомогою методу найменших квадратів побудувати статичну модель, заданої таблицею даних варіанту (наведені нижче), лінійною функцією $y = b_0 + b_1 * x_1 + \cdots + b_n * x_n$;
5. Оцінити адекватність моделі (середньоквадратична похибка, критерій Фішера) та значущість її параметрів.

Завдання за варіантом *(3)*

Під час передвиборчої кампанії кандидат в депутати в кожному з 50 населених пунктів свого округу організовував зустрічі з масами, концерти, роздачу продуктів харчування і друкованих агітаційних матеріалів. У таблиці наведені дані про рейтинг кандидата в кожному населеному пункті і про витрати грошей на кожен захід на 1 жителя пункту. Вважаючи інші умови проведення передвиборної кампанії однаковими, необхідно знайти залежність рейтингу від зазначених факторів.

Дані, отримані під час експерименту:

№	Витрати на зустрічі (гр)	Витрати на концерти (гр)	Витрати на харчування (гр)	Витрати на друк (гр)	Рейтинг (%)
1	4,99	4,99	2,33	8,32	54,17
2	5,22	5,55	2,42	10,42	48,64
3	3,42	5,03	2,19	12,44	45,67
4	4,52	3,90	2,77	10,97	50,51

5	4,31	4,90	2,05	10,11	52,62
6	3,24	3,78	3,60	10,99	51,82
7	4,29	6,11	2,83	12,63	42,61
8	4,04	4,33	3,09	15,89	27,15
9	3,99	3,56	2,94	10,82	52,16
10	1,92	4,57	3,50	8,47	55,20
11	4,66	2,94	2,30	14,67	34,31
12	4,03	2,79	3,52	10,21	54,02
13	5,82	4,32	1,69	10,82	46,11
14	4,66	3,98	2,21	9,91	53,07
15	4,47	4,61	2,11	13,96	37,68
16	3,75	5,07	2,41	12,90	43,46
17	3,52	5,17	4,03	10,04	52,99
18	3,76	2,74	1,10	16,32	25,56
19	2,97	5,01	2,02	5,14	57,22
20	4,81	4,69	3,63	9,58	52,58
21	5,05	5,36	3,32	12,68	41,28
22	5,42	4,22	2,49	11,20	46,72
23	3,17	4,01	1,86	10,18	54,21
24	4,34	3,38	4,26	90,28	-54,9
25	2,46	3,53	1,76	8,11	57,80
26	3,44	2,50	2,11	12,34	47,82
27	5,16	4,34	2,95	10,55	49,52
28	4,16	3,17	,93	10,76	52,22
29	5,28	1,59	3,37	10,59	49,81
30	4,50	3,83	3,03	12,44	45,18
31	5,23	5,56	3,57	8,19	52,70
32	4,76	4,00	2,02	10,58	51,04
33	5,05	5,40	3,75	10,91	47,63
34	5,73	3,84	4,03	7,96	52,18
35	4,40	4,98	3,17	10,57	50,99
36	3,37	3,91	1,66	13,19	43,34
37	3,26	3,15	2,81	-11,30	151,4
38	4,69	4,18	2,07	10,01	52,62
39	4,00	4,48	2,15	14,31	36,72
40	4,04	2,23	3,01	12,61	46,06
41	3,54	4,03	3,07	13,01	43,93
42	4,72	3,17	2,05	8,15	56,45
43	4,43	3,53	3,49	8,02	56,92
44	3,94	5,84	3,16	13,07	41,44
45	4,95	2,52	2,24	13,83	38,18
46	2,42	2,86	2,91	9,18	56,45
47	4,78	4,80	2,24	12,51	43,53
48	2,32	4,18	1,80	14,66	34,73
49	4,59	3,06	2,90	14,17	37,16
50	3,72	3,46	2,84	13,86	40,10

Хід роботи

1. Перевірити дані на наявність помилок та виключити їх (обґрунтувати застосування методів виключення).

Спочатку, дані було перенесено до таблиці в Microsoft Excel:

	A	B	C	D	E	F	G
1	№	Витрати на зустрічі (гр)	Витрати на концерти (гр)	Витрати на харчування (гр)	Витрати на друк (гр)	Рейтинг (%)	
2	1	4.99	4.99	2.33	8.32	54.17	
3	2	5.22	5.55	2.42	10.42	48.64	
4	3	3.42	5.03	2.19	12.44	45.67	
5	4	4.52	3.90	2.77	10.97	50.51	
6	5	4.31	4.90	2.05	10.11	52.62	
7	6	3.24	3.78	3.60	10.99	51.82	
8	7	4.29	6.11	2.83	12.63	42.61	
9	8	4.04	4.33	3.09	15.89	27.15	
10	9	3.99	3.56	2.94	10.82	52.16	
11	10	1.92	4.57	3.50	8.47	55.20	
12	11	4.66	2.94	2.30	14.67	34.31	

З метою дослідження об’єкта дані експерименту піддаються статистичній обробці та аналізу. На основі даних було розраховане середнє значення, середньоквадратичне відхилення, мінімум, максимум кожного зі стовпчиків з даними:

	Витрати на зустрічі (гр)	Витрати на концерти (гр)	Витрати на харчування (гр)	Витрати на друк (гр)	Рейтинг (%)
Середнє:	4.19	4.06	2.68	12.42	47.22
Стандартне відхилення з генеральної сукупності:	0.87990429	0.993859648	0.751199681	11.7824923	21.94624
Стандартне відхилення з вибірки:	0.888837558	1.003949852	0.758826269	11.90211458	22.16905
Мінімальне:	1.92	1.59	0.93	-11.30	-54.90
Максимальне:	5.82	6.11	4.26	90.28	151.40

Маємо некоректні дані:

1) Від’ємні витрати на друк, рейтинг більше 100%.

№	Витрати на зустрічі (гр)	Витрати на концерти (гр)	Витрати на харчування (гр)	Витрати на друк (гр)	Рейтинг (%)
37	3.26	3.15	2.81	-11.30	151.40

2) Від’ємний рейтинг.

№	Витрати на зустрічі (гр)	Витрати на концерти (гр)	Витрати на харчування (гр)	Витрати на друк (гр)	Рейтинг (%)
24	4.34	3.38	4.26	90.28	-54.90

Перевіряємо розрахунки після видалення вищевказаних рядків. Нестандартних даних більше немає:

	Витрати на зустрічі (гр)	Витрати на концерти (гр)	Витрати на харчування (гр)	Витрати на друк (гр)	Рейтинг (%)
Середнє:	4.20	4.10	2.64	11.30	47.17
Стандартне відхилення з генеральної сукупності:	0.887619099	1.000361653	0.730634445	2.309363329	7.63735
Стандартне відхилення з вибірки:	0.897012155	1.010947786	0.738366242	2.333801717	7.7181707
Мінімальне:	1.92	1.59	0.93	5.14	25.56
Максимальне:	5.82	6.11	4.03	16.32	57.80

2. Побудувати гістограму по кожному стовпцю даних в таблиці.
Витрати на зустрічі (гр):

Інтервал 3-х сигм, min	1.511255201
Інтервал 3-х сигм, max	6.893328133

Витрати на концерти (гр):

Інтервал 3-х сигм, min	1.063406643
Інтервал 3-х сигм, max	7.129093357

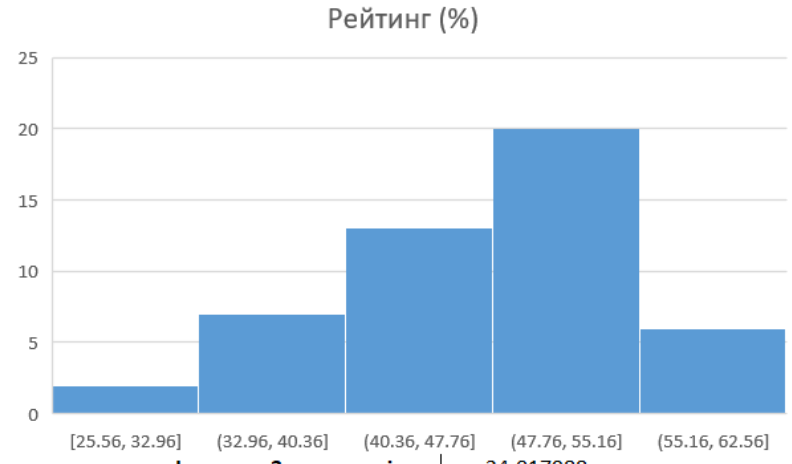
Витрати на харчування (гр):

Інтервал 3-х сигм, min	0.424276274
Інтервал 3-х сигм, max	4.854473726

Витрати на друк (гр):

Інтервал 3-х сигм, min	4.29546985
Інтервал 3-х сигм, max	18.29828015

Рейтинг (%):



Інтервал 3-х сигм, min	24.017988
Інтервал 3-х сигм, max	70.327012

З гістограм по стовпчиках можна зробити висновок, що всі дані входять в потрібний проміжок та задовільняють правило 3-х сигм, тобто практично всі значення нормально розподіленої випадкової величини лежать в інтервалі $[\bar{x} - 3\sigma; \bar{x} + 3\sigma]$.

3. Оцінити взаємний вплив характеристик, наведених в таблиці.

Для оцінки зв'язності факторів між собою та впливу на відгук застосовується побудова кореляційної матриці для даних (після видалення помилок).

Знайдемо коефіцієнт кореляції для кожного з факторів та оцінимо ступінь зв'язку за допомогою шкали Чеддока:

Шкала Чеддока

Значення коефіцієнта кореляції	Зв'язок
[0,1...0,3)	незначний
[0,3...0,5)	помірний
[0,5...0,7)	істотний
[0,7...0,9)	високий
[0,9...0,99]	дуже високий
1,0	функціональний

Бачимо, що коефіцієнт корреляції за витратами на зустрічі, витратами на концерти та витратами на харчування відносно рейтингу низькі, тому їх потрібно видалити. Зв'язок між витратами на друк та рейтингом – дуже високий, майже функціональна залежність.

Матриця кореляції					
	Витрати на зустрічі (гр)	Витрати на концерти (гр)	Витрати на харчування (гр)	Витрати на друк (гр)	Рейтинг (%)
Витрати на зустрічі (гр)		0.086351838	0.150340134	-0.030981724	-0.049583
Витрати на концерти (гр)			0.165543564	-0.136719514	0.0418273
Витрати на харчування (гр)				-0.218670233	0.2157401
Витрати на друк (гр)					-0.932813
Рейтинг (%)					

Таблиця після видалення даних:

	A	B	C
1	№	Витрати на друк (гр)	Рейтинг (%)
2	1	8.32	54.17
3	2	10.42	48.64

4. За допомогою методу найменших квадратів побудувати статичну модель, заданої таблицею даних, лінійною функцією $y = b_0 + b_1 * x_1 + \cdots + b_n * x_n$.

Необхідно знайти невідомі коефіцієнти b. Знаходимо невідомі коефіцієнти многочлена $y = b_0 + b_1 x_1 + \cdots + b_n x_n$ за формулою:

$$B = [U^T * U]^{-1} * U^T * Y$$

Де B – стовпчик невідомих коефіцієнтів, U – матриця даних, що складається із значень факторів, Y – стовпець даних, що містить дані відгуків:

$$U = \begin{pmatrix} 1 & x_{11} & \dots & x_{1n} \\ 1 & \dots & \dots & \dots \\ 1 & x_{m1} & \dots & x_{mn} \end{pmatrix}, Y = \begin{pmatrix} y_1 \\ \dots \\ y_m \end{pmatrix}$$

U^T – транспонована матриця значень факторів.

$[U^T * U]^{-1}$ – обернена матриця добутку матриці значень відгуків на її транспоновану матрицю.

Вигляд матриці U та Y:

U		Y
1	8.32	54.17
1	10.42	48.64
1	12.44	45.67
1	10.97	50.51
1	10.11	52.62
▪▪▪	▪▪▪	▪▪▪

Знаходимо U^T за допомогою функції *TRANSPOSE*:

U^T				
1	1	1	1	▪▪▪
8.32	10.42	12.44	10.97	▪▪▪

Добуток транспонованої матриці на початкову $[U^T * U]$ (функція *MMULT*):

U^T * U	
48	542.25
542.25	6381.722

Знайдемо $[U^T * U]^{-1}$ за допомогою функції *MINVERSE*:

(U^T * U)^-1	
0.519363	-0.04413
-0.04413	0.003906

Розрахунок стовпця параметрів B за формулою МНК ($B = [U^T * U]^{-1} * U^T * Y$):

(U^T * U)^-1 * U^T * Y	
b0	82.02254
b1	-3.08493

Отже,

$$b_0 = 82.02254044$$
$$b_1 = -3.084927508$$

Відгук, обчислений за допомогою знайдених коефіцієнтів:

Витрати на друк (гр)	Рейтинг (%)	Y^M
8.32	54.17	56.35594
10.42	48.64	49.8776
12.44	45.67	43.64604
10.97	50.51	48.18089
10.11	52.62	50.83392
10.99	51.82	48.11919
12.63	42.61	43.05991
15.89	27.15	33.00304
10.82	52.16	48.64362
8.47	55.20	55.8932
14.67	34.31	36.76665
10.21	54.02	50.52543
10.82	46.11	48.64362
9.91	53.07	51.45091
13.96	37.68	38.95695
12.90	43.46	42.22698

5. Оцінити адекватність моделі (середньоквадратична похибка, критерій Фішера) та значущість її параметрів.

Додаємо стовбчик, в якому містяться квадрати різниці відгуку за даними та обчисленого за допомогою знайдених коефіцієнтів, $(Y_i - Y_i^M)^2$.

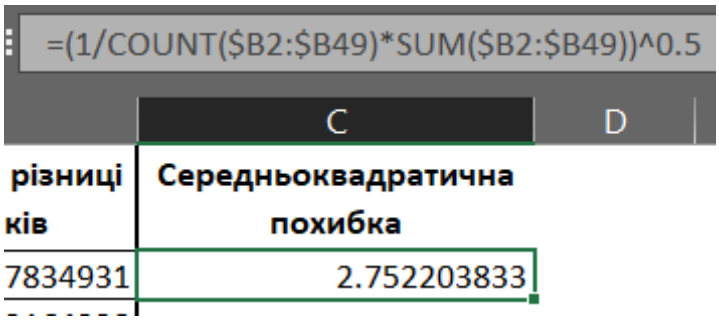
№	Квадрати різниці відгуків
1	4.77834931
2	1.53164338
3	4.09640500
4	5.42477352
5	3.19006985

Знайдемо середньоквадратичну похибку при порівнянні експериментальних даних Y з розрахованими значеннями моделі Y^M.

Формула:

$$\Delta = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(y_i - y_i^M)^2}$$

$m = 48$, кількість експериментів.
Після обчислень отримуємо наступний результат:

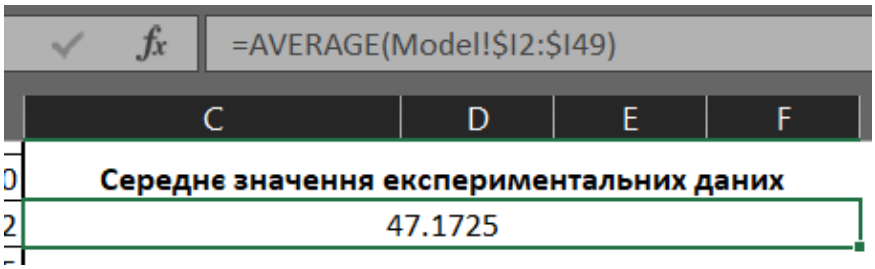


Відповідно, $\Delta = 2.752203833$.

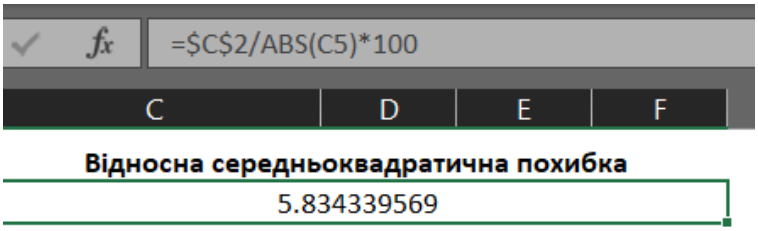
Відносна середньоквадратична похибка розраховується так:

$$\delta = \frac{\Delta}{|y_{\text{сер}}|} 100$$

де $y_{\text{сер}}$ – середнє значення відгуку в експериментальних даних.
Відповідно, $y_{\text{сер}} = 47.1725$:



Відносна середньоквадратична похибка: $\delta = 5.834339569\%$.



Середню відносну похибку можна розрахувати по формулі:

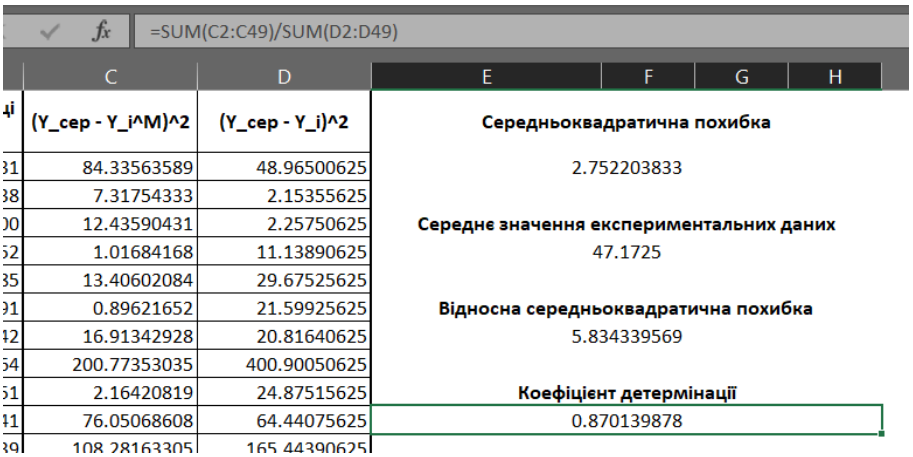
$$\delta = \frac{1}{m} \sum_{i=1}^{m} \frac{|y_i - y_i^M|}{|y_i|} 100$$

Результат: $\delta = 4.78756791\%$

Коефіцієнт детермінації (R^2) показує, наскільки знайдена залежність близька до аналітичного закону (приймає значення від 0 до 1, 1 – аналітичний закон). Для достатньо якісних моделей він має бути більшим від 0,8. Для того, щоб модель вважалася мінімально прийнятною, коефіцієнт має бути більше 0.5. Він розраховується за формулою:

$$R^2 = \frac{\sum_{i=1}^{m}\left(y_{\text{сер}} - y_i^M\right)^2}{\sum_{i=1}^{m}\left(y_{\text{сер}} - y_i\right)^2}$$

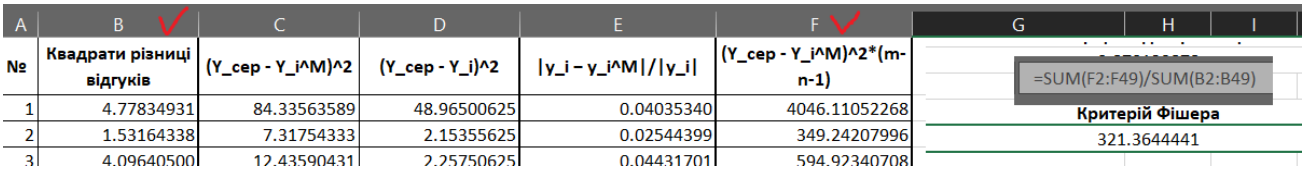
За результатами розрахунків, $R^2 = 0.870139878$.



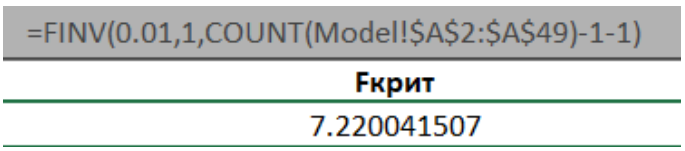
Критерій Фішера розраховується по формулі:

$$F = \frac{\sum_{i=1}^{m}\left(y_{\text{сер}} - y_i^M\right)^2 (m-n-1)}{\sum_{i=1}^{m}(y_i - y_i^M)^2 n}$$

де n – кількість факторів в моделі; m – кількість експериментів.
Значення n та $(m-n-1)$ називають ступенями вільності моделі.
Результат: $F = 321.3644441$



Знаходимо табличне значення критерію Фішера за допомогою функції *FINV*:



Розраховане значення критерію Фішера набагато більше ніж $F_{\text{крит}} = 7.220041507$
Модель можна вважати адекватною.

Для розробленої моделі перевірку адекватності краще проводити на основі нової серії експериментів.

Для кожного параметра моделі $b_0, \dots, b_n$ розраховується коефіцієнт значущості на основі t-критерію Стьюдента по формулам:

$$t_0 = |b_0| \frac{\sqrt{(m-n-1)}}{\sigma}$$

$$t_i = |b_i| \frac{\sqrt{(m-n-1)}}{\sigma_0} \sigma_i, i = \overrightarrow{1,n}$$

де $\sigma_0 = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(y_j - y_j^M\right)^2}$,

$$\sigma_i = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(x_j^i - x_{\text{сер}}^i\right)^2}$$

$x_{\text{сер}}^i$ – середнє значення фактору.

Були пораховані значення σ. Відповідно, $\sigma_0 = 2.752203833,\ \sigma_1 = 2.309363329.$

Квадрати різниці відгуків	sigma_0	(x1 - x_cep)^2	sigma_1
4.77834931	2.752204	8.86178477	2.309363
1.53164338		0.76890977	

Коефіцієнти значущості на основі t-критерію Стьюдента:

=ABS($J4)*(COUNT($A$2:$A$49)-1-1)^0.5/$F$2

7		b0	82.02254
7		b1	-3.08493
7			
7		t_0	202.1304
7		t_1	17.5564
7			

Отже, $t_0 = 202.1304$, $t_1 = 17.5564$.

Також, було знайдене табличне значення $t_{\text{крит}}$ за допомогою функції *T.INV.2T*:

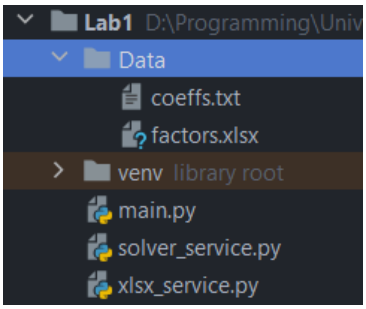
t_0	202.1304
t_1	17.5564

Ступінь вільності	46
t_крит	2.012896

Отже, $t_{крит} = 2.012896$. Всі коефіцієнти значущості більші, ніж це значення, тобто параметри моделі є статистично значущими.

Модель є адекватною за критерієм Фішера і всі її параметри є статистично значущими. З цього випливає, що модель можна використовувати на практиці.

Було створено програму, яка може обчислити значення відгуків, якщо на вхід йде таблиця зі заданими факторами та параметри створеної моделі. Програма була розроблена використовуючи мову програмування Python.



xlsx_service – набір методів, які пов'язані зі зчитуванням з таблиці та записом в неї результатів.

```
import openpyxl


def get_table(filename):
    sheet = openpyxl.load_workbook(filename).active

    table = []
    for row in sheet.iter_rows(values_only=True):
        row_data = []

        # Process the data in the row
        for cell_value in row:
            row_data.append(cell_value)

        table.append(row_data)

    return table


def write_table(filename, model_values):
    workbook = openpyxl.Workbook()
    worksheet = workbook.active

    for i, number in enumerate(model_values):
        worksheet.cell(row=i + 1, column=1, value=number)

    workbook.save(filename)
```

solver_service – набір методів, які пов'язані з валідацією вхідних даних та розрахунками.

```
def validate(table):
    for row in table:
        for value in row:
```

```
            if value is None:
                raise ArithmeticError("Columns have different amount of numbers")


def solve(coeffs, table):
    model_values = []

    for row in table:
        value = coeffs[0]
        i = 1

        for cell_value in row:
            value += coeffs[i] * cell_value
            i += 1

        model_values.append(value)

    return model_values
```

main – точка запуску програми. Використовується для отримання даних та виклику основних функцій.

```
import sys

import xlsx_service
import solver_service


def main():
    if len(sys.argv) != 2:
        raise ValueError("There must be 1 argument: path to .xlsx file")

    excel_file_name = sys.argv[1]

    table = xlsx_service.get_table(excel_file_name)
    solver_service.validate(table)

    coeffs = []
    for i in range(0, len(table[0]) + 1):
        coef = float(input(f"b{i}: "))
        coeffs.append(coef)

    model_values = solver_service.solve(coeffs, table)
    xlsx_service.write_table(".\\Data\\results.xlsx", model_values)
    print("Results were written successfully")


if __name__ == '__main__':
    main()
```

Вкажемо параметри створеної моделі в програмі:

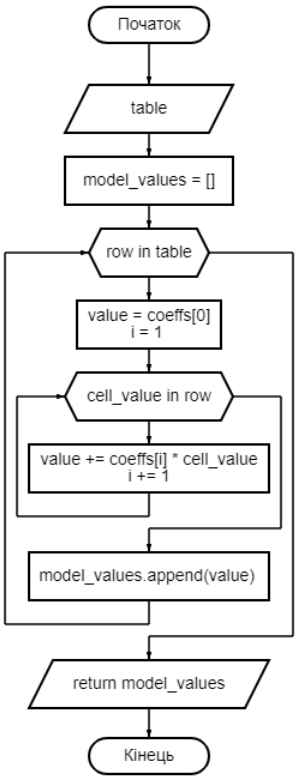
```
b0: 82.02254044
b1: -3.084927508
Results were written successfully
```

Перевіримо створену таблицю:

	A	Y^M
1	56.35594	56.35594
2	49.8776	49.8776
3	43.64604	43.64604
4	48.18089	48.18089
5	50.83392	50.83392
6	48.11919	48.11919
7	43.05991	43.05991
8	33.00304	33.00304
9	48.64362	48.64362
10	55.8932	55.8932
11	36.76665	36.76665
12	50.52543	50.52543

Результати майже ідентичні.

Блок-схема:



Висновок: Під час виконання лабораторної роботи була опрацьована задана за варіантом таблиця, побудовані гістограми, оцінений взаємний вплив характеристик, побудувана статична модель, та оцінена адекватність моделі (середньоквадратична похибка, критерій Фішера) та значущість її параметрів. Також створений програмний додаток для отримання відгуку за факторами та вхідними параметрами моделі.