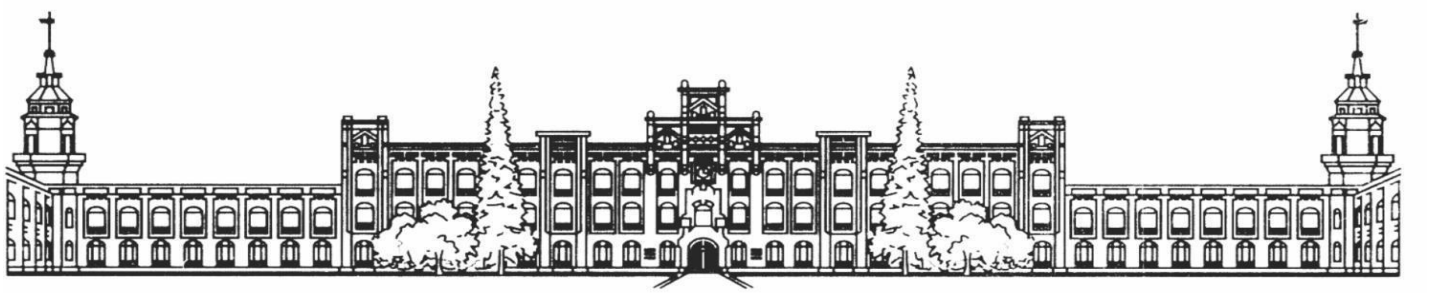


Міністерство освіти і науки України

Національний технічний університет України

«Київський політехнічний інститут імені Ігоря Сікорського»

Інститут прикладного системного аналізу

Розрахункова робота

з регресійного аналізу

Виконав:

студент 2 курсу групи КА-02

Козак Назар Ігорович

Перевірила:

Каніовська І. Ю.

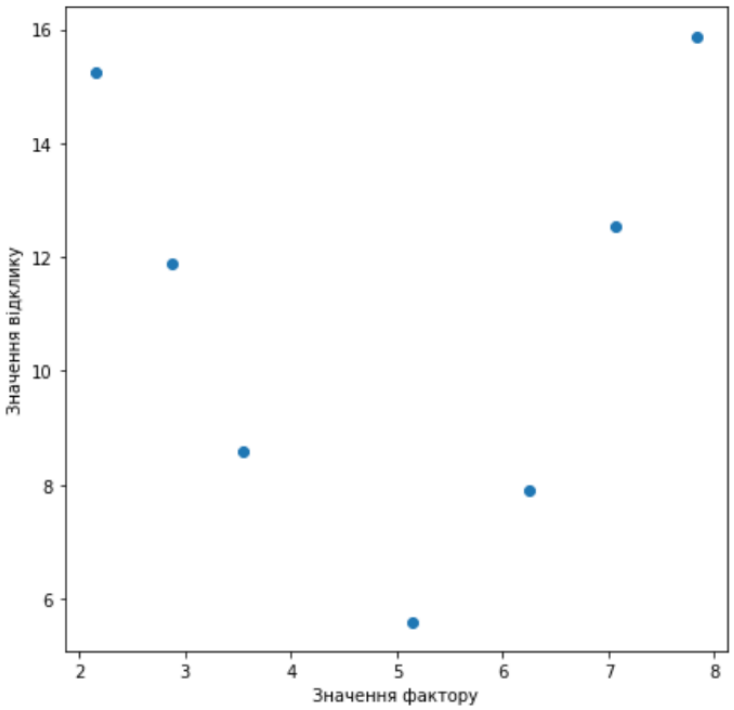
Київ 2022

1 Завдання 1.

1. Провести аналіз вибірки та вибрати підходящу лінійну регресійну модель.
2. За методом найменших квадратів знайти оцінки параметрів вибраної моделі.
3. На рівні значущості $\alpha = 0.05$ перевірити адекватність побудованої моделі.
4. Для найменшого значення параметра побудованої моделі на рівні значущості $\alpha = 0.05$ перевірити гіпотезу про його значущість.
5. Побудувати прогнозований довірчий інтервал з довірчою ймовірністю $g = 0.95$ для середнього значення відклику та самого значення відклику в деякій точці, яку треба обрати самому.
6. Написати висновки.

X	2.15	2.87	3.55	5.14	6.25	7.07	7.83
Y	15.24	11.9	8.6	5.6	7.9	12.54	15.88

1.1 Провести аналіз вибірки та вибрати підходящу лінійну регресійну модель.



За розташуванням точок на діаграмі розсіювання, бачимо що точки на площині розташовані не лінійно, а більше нагадують параболу. Розглянемо модель такого вигляду:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

Мною вибрана лійнійна регресійна модель з такими базисними функціями: $\{1, x, x^2\}$. Для підвищення "точності"моделі можна було б розглядати поліноміальну модель з більшим макимальним сетепенем. В такому випадку модель проходила би ближче до точок зображених на діаграмі розсіювання, але таке ускладнення моделі може призвести до перенавчання(англійською - overfitting). Цей термін означає, що модель на нових значеннях факторів буде погано оцінювати функцію $f(x) = \mathbb{E}(\eta/\xi = x_i)$

Наглядно перенавчання продемонстровано на рисунках нище. Там червні точки це точки "значення фактору - значення відклику". Сині лінії це поліноміальні регресійні моделі з максимальним степенем - M.

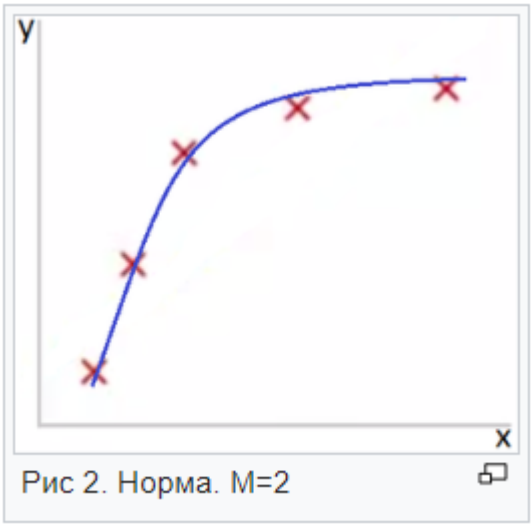


Рис 2. Норма. M=2

Рис 3. Переобучение. M=4

У випадку, якщо вибрана модель не пройде перевірку на адекватність, то виберемо іншу, яка має більший степінь.

1.2 За методом найменших квадратів знайти оцінки параметрів вибраної моделі.

Знайдемо матрицю плану для вибраної моделі:

$$F = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2.15 & 2.87 & 3.55 & 5.14 & 6.25 & 7.07 & 7.83 \\ 2.15^2 & 2.87^2 & 3.55^2 & 5.14^2 & 6.25^2 & 7.07^2 & 7.83^2 \end{pmatrix}^T$$

Оскільки $rangF = 3$, то для того, щоб ми могли використовувати метод найменших квадратів треба зробити припущеня лише про розподіл вектора похибок спостережень(а саме $\vec{\varepsilon} \sim N(\vec{0}, \sigma^2 I)$, де I – одинична матриця).

Тепер знайдемо інформаційну матрицю A і дисперсійну матрицю Фішера A^{-1}:

$$A = F^T F = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 2.15 & 2.87 & \cdots & 7.83 \\ 2.15^2 & 2.87^2 & \cdots & 7.83^2 \end{pmatrix} \cdot \begin{pmatrix} 1 & 2.15 & 2.15^2 \\ 1 & 2.87 & 2.87^2 \\ \vdots & \vdots & \vdots \\ 1 & 7.83 & 7.83^2 \end{pmatrix} \approx \begin{pmatrix} 7 & 34.86 & 202.238 \\ 34.86 & 202.238 & 1291.7 \\ 202.238 & 1291.7 & 8729.18 \end{pmatrix}$$

$$A^{-1} \approx \begin{pmatrix} 7 & 34.86 & 202.238 \\ 34.86 & 202.238 & 1291.7 \\ 202.238 & 1291.7 & 8729.18 \end{pmatrix}^{-1} \approx \begin{pmatrix} 8.2417 & -3.6636 & 0.3512 \\ -3.6636 & 1.7186 & -0.1694 \\ 0.3512 & -0.1694 & 0.0171 \end{pmatrix}$$

Перевіримо властивості інформаційної матриці A:

1. Оскільки F - матриця 7 × 3, а F^T - матриця 3 ×7, то матриця $A = F^T F$ має мати розмірність 3×3. Як бачимо, ця умова виконується

2. A - має бути симетрична. Виконується

3. A - має бути додатньо визначена. Первевіримо це за критерієм Сильвестра:

$$\Delta_1 = 7 > 0$$

$$\Delta_2 = \begin{vmatrix} 7 & 34.86 \\ 34.86 & 202.238 \end{vmatrix} \approx 200.44 > 0$$

$$\Delta_3 = \begin{vmatrix} 7 & 34.86 & 202.238 \\ 34.86 & 202.238 & 1291.7 \\ 202.238 & 1291.7 & 8729.18 \end{vmatrix} \approx 11747.17 > 0$$

Отже, матриця A - додатньо визначена

Тепер враховуючи те, що вектор значень відкликів дорівнює: $\vec{\eta_{\text{зн}}} = (15.24; 11.9; \dots; 15.88)^T$, можемо за формулою $\vec{\beta^*_{\text{зн}}} = A^{-1}F^T\vec{\eta_{\text{зн}}}$ знайти значення оцінок параметрів нашої моделі:

$$A^{-1}F^T \approx \begin{pmatrix} 8.2417 & -3.6636 & 0.3512 \\ -3.6636 & 1.7186 & -0.1694 \\ 0.3512 & -0.1694 & 0.0171 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 2.15 & 2.87 & \cdots & 7.83 \\ 2.15^2 & 2.87^2 & \cdots & 7.83^2 \end{pmatrix} \approx$$

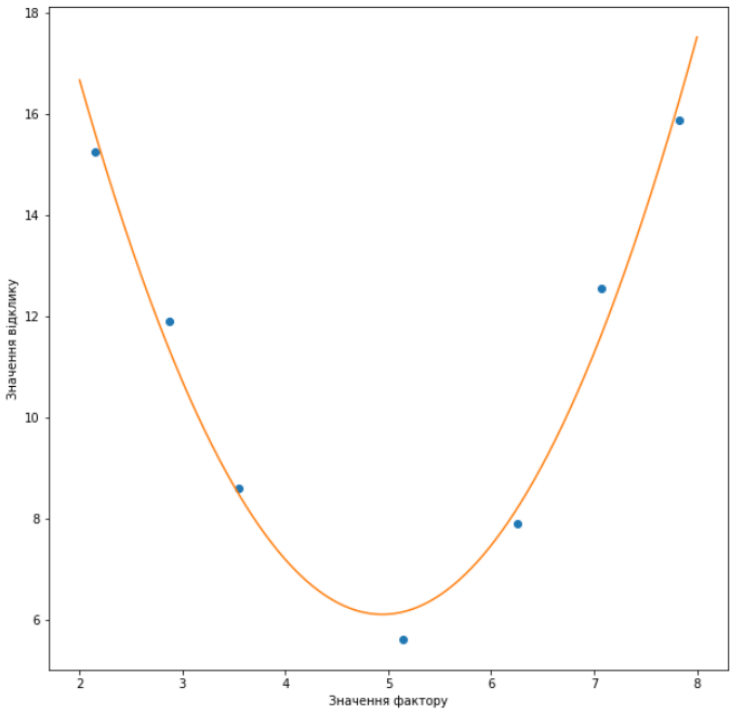
$$\approx \begin{pmatrix} 1.9883 & 0.6198 & -0.3383 & -1.3112 & -0.938 & -0.1065 & 1.0859 \\ -0.7518 & -0.1268 & 0.3022 & 0.6937 & 0.4593 & 0.0179 & -0.5946 \\ 0.0657 & 0.0053 & -0.0354 & -0.0692 & -0.0418 & 0.0055 & 0.0698 \end{pmatrix}$$

$$\vec{\beta^*_{\text{зн}}} = A^{-1}F^T\vec{\eta_{\text{зн}}} \approx \begin{pmatrix} 1.9883 & 0.6198 & \cdots & 1.0859 \\ -0.7518 & -0.1268 & \cdots & -0.5946 \\ 0.0657 & 0.0053 & \cdots & 0.0698 \end{pmatrix} \cdot \begin{pmatrix} 15.24 \\ 11.9 \\ \vdots \\ 15.88 \end{pmatrix} \approx \begin{pmatrix} 35.9238 \\ -12.071 \\ 1.22124 \end{pmatrix}$$

Отримали таку модель:

$$f^*_{\text{зн}}(x) = 35.9238 - 12.071x + 1.22124x^2$$

Зобразимо графік отриманої моделі на діаграмі розсіювання.



1.3 На рівні значущості $\alpha = 0.05$ перевірити адекватність побудованої моделі.

Для перевірки моделі на адекватність скористаємось F-критерієм. Він перевіряє чи є побудована модель кращою за найпростішу - константну. Висунемо основну гіпотезу: H_0 : константна модель та побудована не відрізняються. Тобто основна гіпотеза означає, що дисперсії похибок цих моделей однакові. Висуваємо також альтернативну гіпотезу H_1 : побудована модель є кращою за константну. Розглянемо статистику:

$$\zeta = \frac{\frac{1}{n-1}\sum\limits_{k=1}^{n}(\eta_k - \bar{\eta})^2}{\frac{1}{n-m}\|\vec{\eta} - F\vec{\beta^*}\|^2} = \frac{\frac{1}{n-1}\sum\limits_{k=1}^{n}(\eta_k - \bar{\eta})^2}{\frac{1}{n-m}\sum\limits_{k=1}^{n}\left(\eta_k - f^*(\vec{x^{(k)}})\right)^2} \sim F(n-1, n-m),$$

де n - кількість спостережень, а m - кількість невідомих параметрів. В нашому випадку $n = 7, m = 3$.

Критична область є правосторонньою: при $\zeta_{\text{зн}} > t_{\text{кр}}$ основна гіпотеза відхиляється і модель вважається адекватною.

Знайдемо значення статистики($\zeta_{\text{зн}}$):

$$(\bar{\eta})_{\text{зн}} = \frac{1}{7}(15.24 + 11.9 + 8.6 + 5.6 + 7.9 + 12.54 + 15.88) \approx 11.0943$$

$$\zeta_{\text{зн}} = \frac{2}{3} \cdot \frac{(15.24 - 11.0943)^2 + \cdots + (15.88 - 11.0943)^2}{(15.24 - 15.6168)^2 + \cdots + (15.88 - 16.2824)^2} \approx 32.2557$$

За таблицею квантилів рівня 0.95 для розподілу Фішера-Снедекора знаходимо значення $t_{\text{кр}}$: оскільки $n_1 = 6, n_2 = 4, \alpha = 0.05$, маємо $t_{\text{кр}} = 6.16$. Оскільки критична область є правостороньою і $\zeta_{\text{зн}} > t_{\text{кр}}$, то на рівні значущості $\alpha = 0.05$ модель можна вважати адекватною. Оскільки модель адекватна, то її не треба замінювати на ту, яка має більший максимальний степінь. Але я вирішив побудувати ще одну, точішу, з використанням мови програмування python та бібліотек numpy і matplotlib. Вона знаходиться в додатку А(після 12 сторінки).

1.4 Для найменшого значення параметра побудованої моделі на рівні значущості $\alpha = 0.05$ перевірити гіпотезу про його значущість.

На рівні значущості $\alpha = 0.05$ перевіримо гіпотезу про значущість параметру $\beta_3\Big((\beta_3^*)_{\text{зн}} = 1.122124\Big)$. Основною гіпотезою є $H_0 : \beta_3 = 0$, альтернативною – $H_1 : \beta_3 > 0$. Критична область є правосторонньою. Розглядаємо статистику:

$$\gamma = \frac{\beta_j^*}{\sqrt{(\sigma^2)^{**} \cdot a_{jj}}} \sim St_{n-m}$$

В нашому випадку $j = 3, n = 7, m = 3$, тому

$$\gamma = \frac{\beta_3^*}{\sqrt{(\sigma^2)^{**} \cdot a_{33}}} \sim St_4$$

Знайдемо значення $\gamma_{\text{зн}}$:

$$((\sigma^2)^{**})_{\text{зн}} = \frac{1}{4}\left\|\vec{\eta}_{\text{зн}} - F(\vec{\beta^*})_{\text{зн}}\right\|^2 \approx 0.161934$$

$$\gamma_{\text{зн}} = \frac{1.122124}{\sqrt{0.161934 \cdot 0.0698}} \approx 11.4869$$

За таблицею деяких квантилів розподілу St_n знаходимо значення $t_{\text{кр}}$. В нашому випадку $\alpha = 0.05, n = 4$, тому $t_{\text{кр}} = 2.132$. Оскільки критична область - правостороння і $\gamma_{\text{зн}} > t_{\text{кр}}$, то ми потрапляємо в критичну область. Основна гіпотеза відхиляється, тому параметр β_3 є значущим.

1.5 Побудувати прогнозований довірчий інтервал з довірчою ймовірністю $g = 0.95$ для середнього значення відклику та самого значення відклику в деякій точці, яку треба обрати самому.

Виберемо точку $x_0 = 6.5$. Під $\vec{x}$ будемо вважати вже вибраний набір значень факторів $\vec{x} = (1, 6.5, 6.5^2)^T$

Для побудови довірчого інтервалу для середнього значення відклику використаємо статистику:

$$\nu = \frac{f^*(\vec{x}) - f(\vec{x})}{\sqrt{(\sigma^2)^{**}\vec{x}^T A^{-1}\vec{x}}} \sim St_{n-m} = St_4$$

Довірчий інтервал для середнього значення відклику має вигляд:

$$f(\vec{x}) \in \left(f^*(\vec{x}) - t\sqrt{(\sigma^2)^{**}\vec{x}^T A^{-1}\vec{x}}, f^*(\vec{x}) + t\sqrt{(\sigma^2)^{**}\vec{x}^T A^{-1}\vec{x}}\right)$$

Обчислимо значення $\vec{x}^T A^{-1}\vec{x}$ та $f^*_{\text{зн}}(\vec{x})$:

$$\vec{x}^T A^{-1}\vec{x} \approx \left(1, 6.5, 6.5^2\right) \begin{pmatrix} 8.2417 & -3.6636 & 0.3512 \\ -3.6636 & 1.7186 & -0.1694 \\ 0.3512 & -0.1694 & 0.0171 \end{pmatrix} \begin{pmatrix} 1 \\ 6.5 \\ 6.5^2 \end{pmatrix} \approx 0.2755$$

$$f^*_{\text{зн}}(\vec{x}) \approx 35.9238 - 12.071x_0 + 1.22124x_0^2 \approx 9.0611$$

За таблицею значень квантилів розподілу Стьюдента знаходимо значення $t = t_{0.025,4} = 2.776$. Отже, піставивши значення маємо:

$$\left(9.0611 - 2.776\sqrt{0.161934 \cdot 0.2755}, 9.0611 + 2.776\sqrt{0.161934 \cdot 0.2755}\right) \approx (8.475, 9.65)$$

Отже, отримали довірчий інтервал для середнього значення відклику у точці $x_0 : f(\vec{x}) \in (8.475, 9.65)$ з ймовірністю 0.95.

Тепер побудуємо прогнозований інтервал з довірчою ймовірністю 0.95 для самого значення відклику. Розглянемо таку статистику:

$$\epsilon = \frac{\eta - f^*(\vec{x})}{\sqrt{(\sigma^2)^{**}(1 + \vec{x}^T A^{-1}\vec{x})}} \sim St_{n-m} = St_4$$

Довірчий інтервал для самого значення відклику має вигляд:

$$\eta \in \left(f^*(\vec{x}) - t\sqrt{(\sigma^2)^{**}(1 + \vec{x}^T A^{-1}\vec{x})}, f^*(\vec{x}) + t\sqrt{(\sigma^2)^{**}(1 + \vec{x}^T A^{-1}\vec{x})}\right)$$

Підставимо значення:

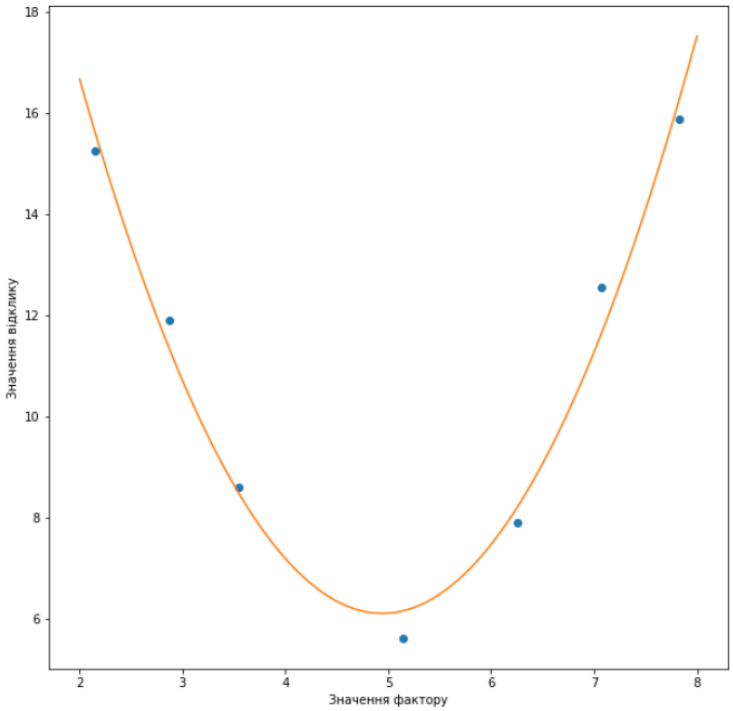
$$\left(9.0611 - 2.776\sqrt{0.161934 \cdot (1 + 0.2755)}, 9.0611 + 2.776\sqrt{0.161934 \cdot (1 + 0.2755)}\right) \approx (7.8, 10.32)$$

Отже, отримали довірчий інтервал для значення відклику у точці $x_0 : \eta \in (7.8, 10.32)$ з ймовірністю 0.95.

1.6 Висновки

Під час виконання першого завдання було проаналізовано вибірку. Оскільки точки на діаграмі розсіювання нагадували параболу було вибрано лінійну регресійну модель такого виду: $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$. Методом найменших квадратів було знайдено значення оцінок параметрів моделі. На рівні значущості 0.05 була перевірена адекватність моделі. Було показано, що на рівні значущості $\alpha = 0.05$ модель можна

вважати адекватною. Згодом був вибраний параметр найменший по модулю і була проведена перевірка його на значущість. Параметр виявся значущим. Було побудовано прогнозовані довірчі інтервали з довірчими ймовірностями рівними 0.95 для середнього значення відклику та самого значення відклику в точці $x_0 = 6.5$ ($f(\vec{x}) \in (8.475, 9.65)$ $\eta_{\text{зн}} \in (7.8, 10.32)$). Модель на діаграмі розсіювання:



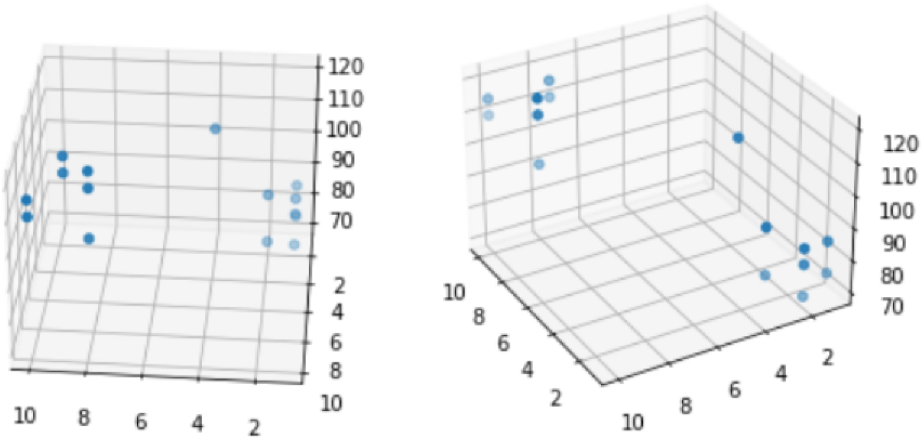
2 Завдання 2.

Дана таблиця експериментальних даних. Треба:

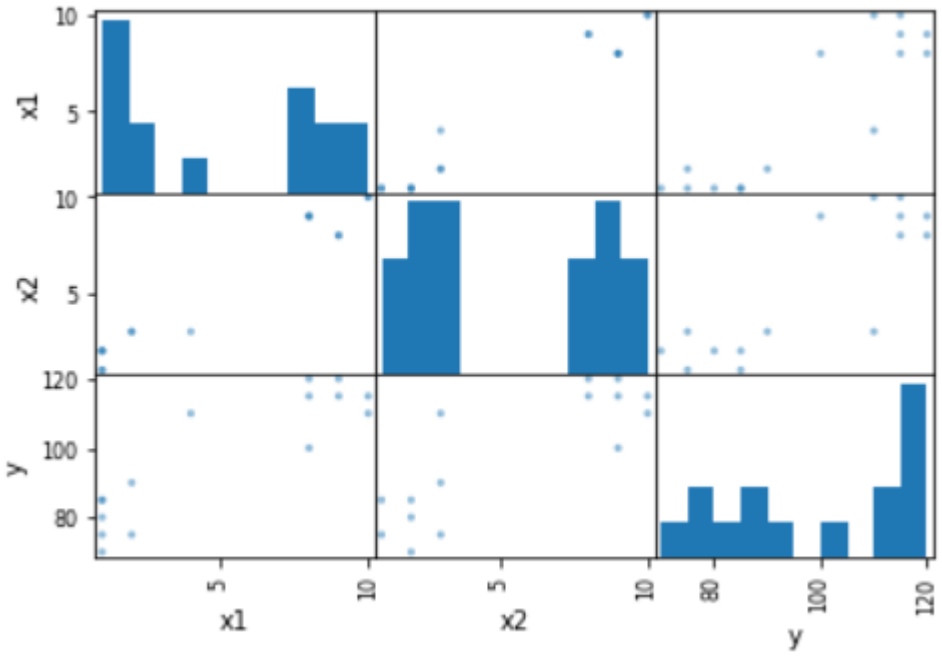
1. За методом найменших квадратів знайти оцінки параметрів двофакторної регресійної моделі.
2. На рівні значушості $\alpha = 0.05$ перевірити адекватність побудованої моделі.
3. Для найменшого значення параметра побудованої моделі на рівні значущості $\alpha = 0.05$ перевірити гіпотезу про його значущість.
4. Побудувати прогнозований довірчий інтервал з довірчою ймовірністю $g = 0.95$ для середнього значення відклику та самого значення відклику в деякій точці(точку треба вибрати самому).
5. Написати висновки.

№	X_{1i}	X_{2i}	Y_i
1.	1	1	75
2.	1	1	85
3.	10	10	115
4.	10	10	110
5.	1	2	85
6.	1	2	70
7.	1	2	80
8.	8	9	100
9.	8	9	115
10.	8	9	120
11.	2	3	75
12.	2	3	90
13.	9	8	120
14.	9	8	115
15.	4	3	100

Зобразимо тривимірну діаграму розсіювання з різних ракурсів:



Візуалізуємо таблицю експериментальних данних за допомогою бібліотеки мови пайтон - Pandas. Побудуємо так звану scatter_marix. Це "матриця"елементами якої є діаграми розсіювання. Ця матриця корисна, оскільки допомагає візуалізувати зв'язок змінними в наборі данних.



Як бачимо, по мірі зростання $x1$ чи $x2$ зростає y. Видно хоч і слабку, але все таки лінійну залежність. Тому розглянемо просту лінійну двофакторну регресійну модель:

$$f(\vec{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

2.1 За методом найменших квадратів знайти оцінки параметрів двофакторної регресійної моделі.

Матриця плану для вибрано моделі має вигляд:

$$F = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 10 & \cdots & 4 \\ 1 & 1 & 10 & \cdots & 3 \end{pmatrix}^T$$

Знайдемо інформаційну матрицю $A = F^T F$, а також дисперсійну матрицю Фішера A^{-1}:

$$A = F^T F = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 10 & 10 \\ \vdots & \vdots & \vdots \\ 1 & 4 & 3 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 10 & \cdots & 4 \\ 1 & 1 & 10 & \cdots & 3 \end{pmatrix} = \begin{pmatrix} 15 & 75 & 80 \\ 75 & 583 & 592 \\ 80 & 592 & 612 \end{pmatrix}$$

$$A^{-1} \approx \begin{pmatrix} 0.2505 & 0.0578 & -0.0886 \\ 0.0578 & 0.11 & -0.1139 \\ -0.0886 & -0.1139 & 0.1234 \end{pmatrix}$$

Перевіримо деякі властивості інформаційної матриці

1. Інформаційна матриця симетрична - виконується.
2. F^T - матриця 3×15, F - матриця 15×3, тому матриця $A = F^T F$ має мати розмірність 3×3 - виконується
3. А має бути додатньо визначеною

$$\Delta_1 = 15 > 0$$
$$\Delta_2 = \begin{vmatrix} 15 & 75 \\ 75 & 583 \end{vmatrix} = 3120 > 0$$
$$\Delta_3 = \begin{vmatrix} 15 & 75 & 80 \\ 75 & 583 & 592 \\ 80 & 592 & 612 \end{vmatrix} = 25280 > 0$$

Отже, за критерієм Сильвестра: матриця А - додатньо визначена

Вектор значень відкликів має вигляд:

$$\vec{\eta_{\text{зн}}} = (75, 85, 115, 110, \ldots, 100)^T$$

Як і в першому завданні, оскільки $rang F = 3$, то, щоб використовувати МНК треба зробити припущення лише про те що вектор похибок спостережень розподілений так: $\vec{\varepsilon} \sim N(\vec{0}, \sigma^2 I)$. Тепер можемо знайти значення оцінок параметрів нашої моделі:

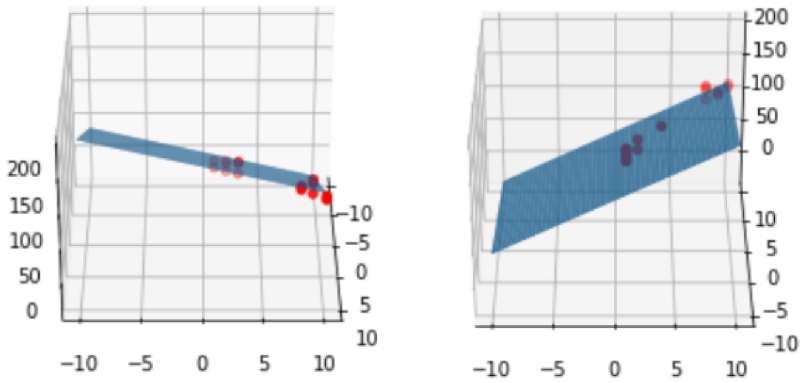
$$\vec{\beta^*_{\text{зн}}} = A^{-1} F^T \vec{\eta_{\text{зн}}} = \begin{pmatrix} 0.2505 & 0.0578 & -0.0886 \\ 0.0578 & 0.11 & -0.1139 \\ -0.0886 & -0.1139 & 0.1234 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 10 & \cdots & 4 \\ 1 & 1 & 10 & \cdots & 3 \end{pmatrix} \cdot (75, 85, 115, 110, \ldots, 100)^T \approx$$

$$\approx \begin{pmatrix} 79.0736 \\ 7.7017 \\ -3.7342 \end{pmatrix}$$

Отримали таку модель:

$$f^*_{\text{зн}}(\vec{x}) = 79.0736 + 7.7017 x_1 - 3.7342 x_2$$

Зобразимо її графік на тривимірній діаграмі розсіювання.



2.2 На рівні значушості $\alpha = 0.05$ перевірити адекватність побудованої моделі.

Як і в попередній задачі, будемо перевіряти адекватність моделі за F критерієм. Розглядаємо статистику:

$$\zeta = \frac{\frac{1}{n-1}\sum\limits_{k=1}^{n}(\eta_k - \bar{\eta})^2}{\frac{1}{n-m}\sum\limits_{k=1}^{n}\left(\eta_k - f^*(\vec{x^{(k)}})\right)^2} \sim F(n-1, n-m)$$

У нашому випадку $n = 15, m = 3$. Висуваємо основну гіпотезу H_0 : константна модель та побудована не відрізняються, а також альтернативну H_1 : побудована модель є кращою за константну. Знайдемо значення статистики($\zeta_{\text{зн}}$):

$$(\bar{\eta})_{\text{зн}} = \frac{1}{15}\Big(75 + 85 + 115 + 110 + \cdots + 100\Big) \approx 97.667$$

$$\zeta_{\text{зн}} = \frac{\frac{1}{14}\left((75 - 97.667)^2 + \cdots + (100 - 97.667)^2\right)}{\frac{1}{12}((75 - 83.0411)^2 + \cdots + (100 - 98.678)^2)} \approx 4.937$$

З таблиці квантилів рівня 0.95 для розподілу Фішера-Снедекора знаходимо значення $t_{\text{кр}} = t_{14,12} \approx 2.64$. Оскільки критична область правостороння і $\zeta_{\text{зн}} > t_{\text{кр}}$, то основна гіпотеза відхиляється і приймається альтернативна. Тобто, на рівні значущості $\alpha = 0.05$ дані не суперечать адекватності моделі.

2.3 Для найменшого значення параметра побудованої моделі на рівні значущості $\alpha = 0.05$ перевірити гіпотезу про його значущість.

На рівні значущості $\alpha = 0.05$ перевіримо гіпотезу про значущість параметру β_3 ($(\beta_3^*)_{\text{зн}} = -3.7342$). Висуваємо основну гіпотезу H_0 : $\beta_3 = 0$ і альтернативну H_1 : $\beta_3 < 0$.. критична область - лівостороння. Розглядаємо статистику:

$$\gamma = \frac{\beta_j^*}{\sqrt{(\sigma^2)^{**} \cdot a_{jj}}} \sim St_{n-m}$$

Обчислимо значення цієї статистики:

$$\gamma_{\text{зн}} = \frac{-3.7342}{\sqrt{\frac{1}{12} \cdot 67.9038 \cdot 0.00004}} \approx -1.29$$

З таблиці квантилів розподілу Стьюдента знаходимо: $t_{\text{кр}} = -t_{0.05,12} = -1.782$. Оскільки критична область лівостороння і $t_{\text{кр}} < \gamma_{\text{зн}}$, то ми попадаємо в область прийняття гіпотези. Отже, на рівні значущості 0.05 ми приймаємо припущення, що параметр β_3 є незначущим. Таким чином маємо нову модель:

$$(f_2^*(\vec{x}))_{\text{зн}} = 79.0736 + 7.7017x_1$$

Перевіримо її на адекватність. Знайдемо значення статистики ζ.

$$\zeta_{\text{зн}} = \frac{\frac{1}{14}\left((75 - 97.667)^2 + \cdots + (100 - 97.667)^2\right)}{\frac{1}{13}((75 - 86.775)^2 + \cdots + (100 - 109.88)^2)} \approx 0.43$$

Тепер ми оцінюємо не 3, а 2 параметри, тому $m = 2$. Знаходимо значення: $t_{\text{кр}} = t_{14,13} = 2.55$. Оскільки критична область правостороння, то приймається основна гіпотеза, тому модель не є адекватною на рівні значущості 0.05. Тому повертаємось до попередньої моделі:

$$f_{\text{зн}}^*(\vec{x}) = 79.0736 + 7.7017x_1 - 3.7342x_2$$

2.4 Побудувати прогнозований довірчий інтервал з довірчою ймовірністю $g = 0.95$ для середнього значення відклику та самого значення відклику в деякій точці(точку треба вибрати самому).

Виберемо точку $\vec{x_0} = (2, 2)^T$. $\vec{x}$ тут вважатимемо вже вибраним набором значень факторів.

1. Інтервал для середнього значення.

   Розглядаємо статистику:

   $$\nu = \frac{f^*(\vec{x}) - f(\vec{x})}{\sqrt{(\sigma^2)^{**}\vec{x}^T A^{-1}\vec{x}}} \sim St_{12}$$

   Шуканий довірчий інтервал має вигляд:

   $$f(\vec{x}) \in \left(f^*(\vec{x}) - t\sqrt{(\sigma^2)^{**}\vec{x}^T A^{-1}\vec{x}}, f^*(\vec{x}) + t\sqrt{(\sigma^2)^{**}\vec{x}^T A^{-1}\vec{x}}\right)$$

   Обчислимо його межі для наших даних:

   $$\vec{x^T} A^{-1}\vec{x} = \begin{pmatrix}1\\2\\2\end{pmatrix} \cdot \begin{pmatrix}0.2505 & 0.0578 & -0.0886\\0.0578 & 0.11 & -0.1139\\-0.0886 & -0.1139 & 0.1234\end{pmatrix} \cdot (1, 2, 2) \approx 0.1492$$

   $$(\sigma^2)^{**}_{\text{зн}} = \frac{1}{15-3}\left\|\vec{\eta}_{\text{зн}} - F\vec{\beta^*}_{\text{зн}}\right\| \approx 67.9038; \qquad t = St_{0.025,12} = 2.179$$

   тому маємо $f(\vec{x}) \in (80.073, 93.944)$ з ймовірністю 0.95

2. Інтервал для самого значення відклику.

   Розглядаємо статистику:

   $$\epsilon = \frac{\eta - f^*(\vec{x})}{\sqrt{(\sigma^2)^{**}(1 + \vec{x}^T A^{-1}\vec{x})}} \sim St_{n-m} = St_{12}$$

   Шуканий довірчий інтервал має вигляд:

   $$\eta \in \left(f^*(\vec{x}) - t\sqrt{(\sigma^2)^{**}(1 + \vec{x}^T A^{-1}\vec{x})}, f^*(\vec{x}) + t\sqrt{(\sigma^2)^{**}(1 + \vec{x}^T A^{-1}\vec{x})}\right)$$

   Обчислимо його межі для наших даних. Отримуємо, що $\eta \in (67.76, 106.2574)$ з ймовірністю 0.95

2.5 Висновки.

Під час виконання другого завдання була проаналізована таблиця експериментальних даних: а саме побудована тривимірна діаграма розсіювання а також за допомогою бібліотеки pandas мови програмування python була побудована так звана scatter_matrix. Було вирішено вибрати просту лінійну двофакторну регресійну модель: $f(\vec{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. За методом найменших квадратів було знайдено значення оцінок параметрів моделі. Отримали таку модель: $f^*_{\text{зн}}(\vec{x}) = 79.0736 + 7.7017x_1 - 3.7342x_2$ Було показано, що на рівні значущості $\alpha = 0.05$ дані не суперечать адекватності побудованої моделі. Була також проведена перевірка найменшого по модулю параметра на значущість. Виявилося, що він є незначущим. Таким чином ми отримали нову модель: $\left(f^*_2(\vec{x})\right)_{\text{зн}} = 79.0736 + 7.7017x_1$. Але, оскільки вона не пройшла перевірку на адекватність, то ми повернулися до попередньої моделі. В кінці були побудовані довірчі інтервали для середнього значення відклику і самого значеняня відклику в точці $x_0 = (2, 2)$

Використана література:

1. Електоронний конспект лекцій – Каніовська І.Ю.
2. Перенавчання моделі лінійної регресії

Програмне забезпечення:

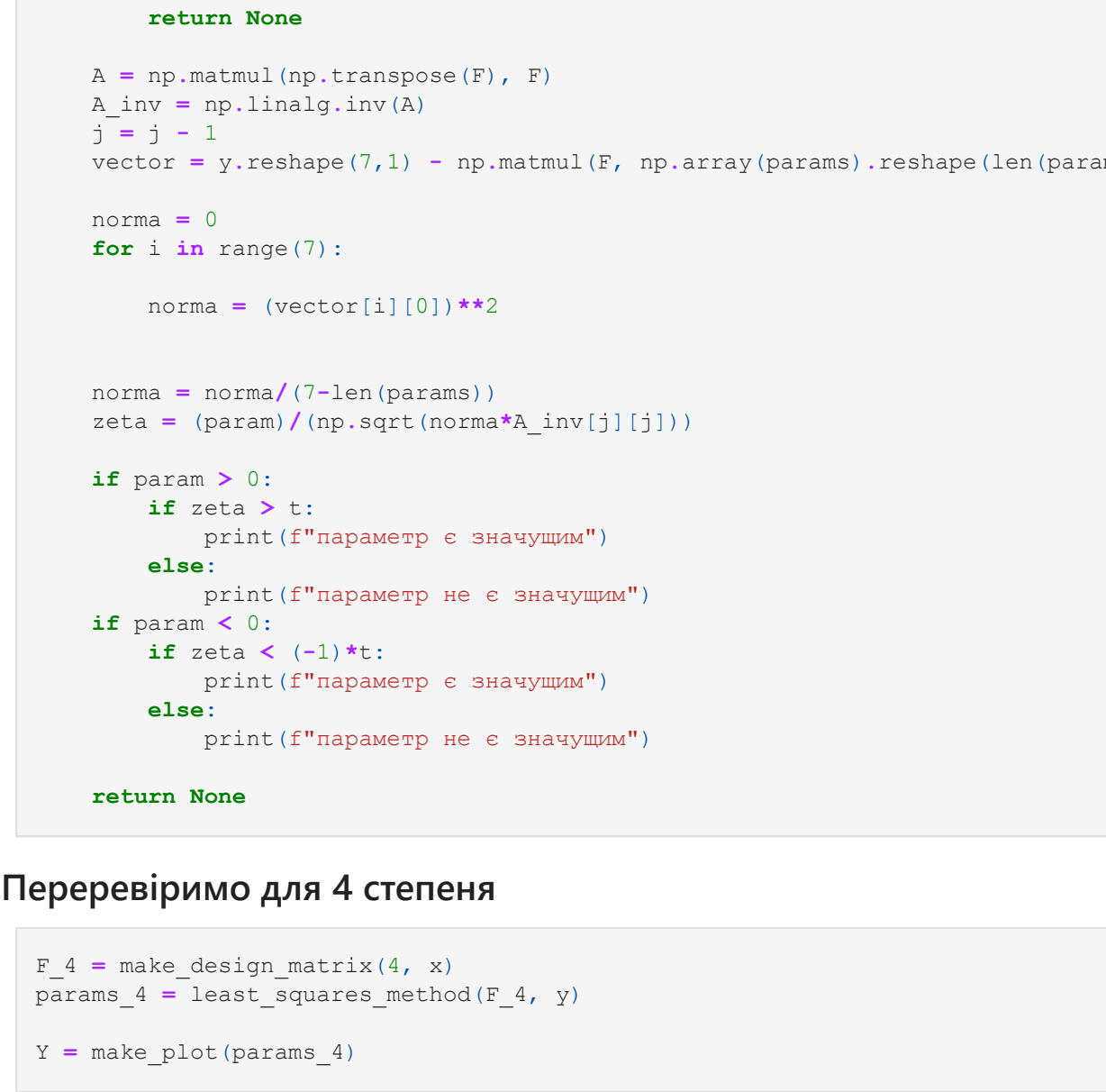
Середовище розробки jupyter notebook. Мова програмування python, а також бібліотеки мови python: pandas, numpy, matplotlib.

Додаток А

In [5]:
```
import matplotlib.pyplot as plt
import numpy as np

x = np.array([2.15, 2.87, 3.55, 5.14, 6.25,  7.07, 7.83])
y = np.array([15.24, 11.9, 8.6, 5.6, 7.9, 12.54, 15.88])

def least_squares_method(F, y):

    A = np.matmul(np.transpose(F), F)
    A_inv = np.linalg.inv(A)
    y = y.reshape(7,1)

    params = np.matmul(A_inv, np.matmul(np.transpose(F), y))

    list = []

    for i in range(params.shape[0]):
        list.append(params[i][0])

    return list


# робить матрицю плану для поліноміальної регресії з макс. степенем k
def make_design_matrix(k, x):


    F = np.array([np.ones(7)])

    for i in range(k):

        F = np.vstack((F, x**(i+1)))

    return np.transpose(F)


def make_plot(params):

    X = np.linspace(2,8,1000)
    Y = 0

    for i in range(len(params)):
        Y = Y + params[i]*(X**(i))

    plt.figure(figsize=(10,10))
    plt.plot([2.15, 2.87, 3.55, 5.14, 6.25,  7.07, 7.83], [15.24, 11.9, 8.6,  5.6, 7.9, 12.54, 15.88], "o")
    plt.plot(X,Y)
    plt.xlabel("Значення фактору")
    plt.ylabel("Значення відклику")

    plt.show()

    return None


#перевірка моделп + і на адекватність
def model_check(x, y, t, f, params):

    if len(params) >= y.size:
        print("неможливо визначити за F-критерієм")
        return None

    bar_y = 0

    for i in range(y.size):
        bar_y = bar_y + y[i]

    bar_y = bar_y/y.size
    n, d = 0, 0

    for i in range(y.size):
        n = n + (y[i] - bar_y)**2
        d = d + (y[i] - f(x[i]))**2

    n = n/(y.size - 1)
    d = d/(y.size - len(params))

    if (n/d) <= t:
        print("модель не є адекватною на рівні значущості 0.05")
    else:
        print("модель є адекватною на рівні значущості 0.05")

    return None

def param_check(F, y, t, param, params, j):

    if len(params) >= y.size:
        print("Помилка: m > n")
        return None

    A = np.matmul(np.transpose(F), F)
    A_inv = np.linalg.inv(A)
    j = j - 1
    vector = y.reshape(7,1) - np.matmul(F, np.array(params).reshape(len(params),1))

    norma = 0
    for i in range(7):

        norma = (vector[i][0])**2


    norma = norma/(7-len(params))
    zeta = (param)/(np.sqrt(norma*A_inv[j][j]))

    if param > 0:
        if zeta > t:
            print(f"параметр є значущим")
        else:
            print(f"параметр не є значущим")
    if param < 0:
        if zeta < (-1)*t:
            print(f"параметр є значущим")
        else:
            print(f"параметр не є значущим")

    return None
```

Перевіримо для 4 степеня

In [2]:
```
F_4 = make_design_matrix(4, x)
params_4 = least_squares_method(F_4, y)

Y = make_plot(params_4)
```



отримали таку модель: $(f_4^*(x))_{\text{зн}} = 6.65318 + 16.98095x - 8.766x^2 + 1.4194x^3 - 0.0712x^4$. Перевіримо її на адекватність

In [3]:
```
def f_4(x):
    return params_4[0] + params_4[1]*x + params_4[2]*(x**2) + params_4[3]*(x**3) + params_4[4]*(x**4)


#у нашому випадку n - 1 = 6; n - m = 2. Тому F(6,2):
t =  6.16

model_check(x, y,t, f_4, params_4)
```

```
модель є адекватною на рівні значущості 0.05
```

На рівні значущості $\alpha = 0.05$ перевіримо гіпотезу про значущість параметру β_5

In [6]:
```
param_check(F_4, y, 2.92, params_4[4], params_4, 5)
```

```
параметр є значущим
```