НАВЧАЛЬНО-НАУКОВИЙ КОМПЛЕКС
"ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ"
НАЦІОНАЛЬНОГО ТЕХНІЧНОГО УНІВЕРСИТЕТУ УКРАЇНИ
"КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО"
КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ СИСТЕМНОГО АНАЛІЗУ

РОЗРАХУНКОВА РОБОТА 2

з предмету "Математична статистика"
з теми "Регресійний аналіз"
Варіант 13

Виконала:
студентка групи
КА-02
Шапошнікова Софія

Перевірила:
Каніовська І.Ю.

Київ 2022

1 Завдання 1

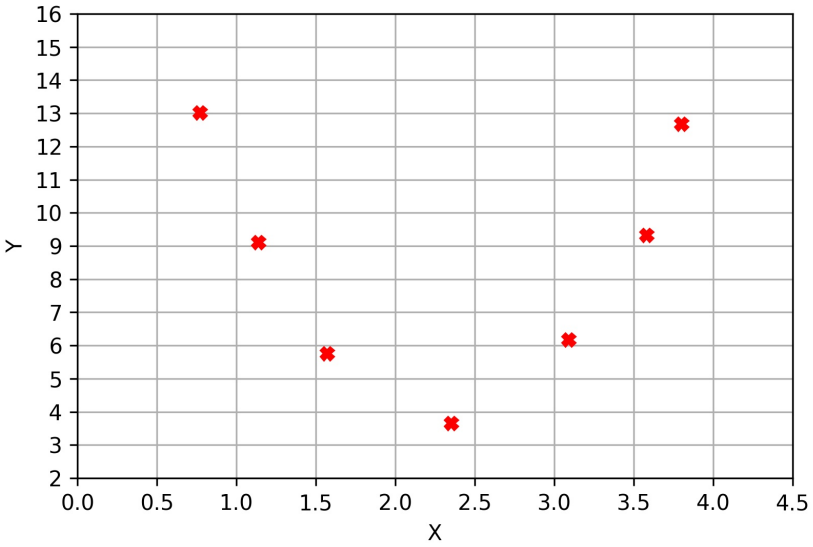
Варіант №13.

X	0,77	1,14	1,57	2,35	3,09	3,58	3,8
Y	13,01	9,11	5,75	3,66	6,17	9,33	12,67

План роботи

1. Провести аналіз вибірки та вибрати підходящу лінійну регресійну модель.
2. За методом найменших квадратів знайти оцінки параметрів вибраної моделі. Перевірити адекватність побудованої моделі.
3. На рівні значущості $\alpha = 0{,}05$ перевірити адекватність побудованої моделі.
4. Для самого малого значення параметра побудованої моделі на рівні значущості $\alpha = 0{,}05$ перевірити гіпотезу про його значущість.
5. Побудувати прогнозований довірчий інтервал з довірчою ймовірністю $\gamma = 0{,}95$ для середнього значення відклику та самого значення відклику в деякій точці (точку вибирайте самі).
6. Написати висновки

1.1 Аналіз вибірки та вибір регресійної моделі



За розташуванням точок – пар значень (фактор-відклик) бачимо, що вони розміщені на площині не лінійно, а більше нагадують параболу - залежність схожа на квадратичну. Розглянемо модель:
$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$

1.2 Знаходження за МНК оцінки параметрів вибраної моделі

Матриця плану для вибраної моделі матиме вигляд:

$$F = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0.77 & 1.14 & 1.57 & 2.35 & 3.09 & 3.58 & 3.8 \\ 0.77^2 & 1.14^2 & 1.57^2 & 2.35^2 & 3.09^2 & 3.58^2 & 3.8^2 \end{pmatrix}^T =$$

$$= \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0.77 & 1.14 & 1.57 & 2.35 & 3.09 & 3.58 & 3.8 \\ 0.5929 & 1.2996 & 2.4649 & 5.5225 & 9.5481 & 12.8164 & 14.44 \end{pmatrix}^T$$

Після цього використовуємо матрицю плану F для знаходження інформаційної матриці Фішера: $A = FF^T$

$$A = \begin{pmatrix} 7 & 16.3 & 46.6844 \\ 16.3 & 46.6844 & 149.044186 \\ 46.6844 & 149.044186 & 502.5541514 \end{pmatrix}^T$$

Знайдемо дисперсійну матрицю Фішера A^{-1}

$$A^{-1} = \begin{pmatrix} 4.24536919 & -4.1988219 & 0.85088836 \\ -4.1988219 & 4.5557084 & -0.96105538 \\ 0.85088836 & -0.96105538 & 0.20797063 \end{pmatrix}^T$$

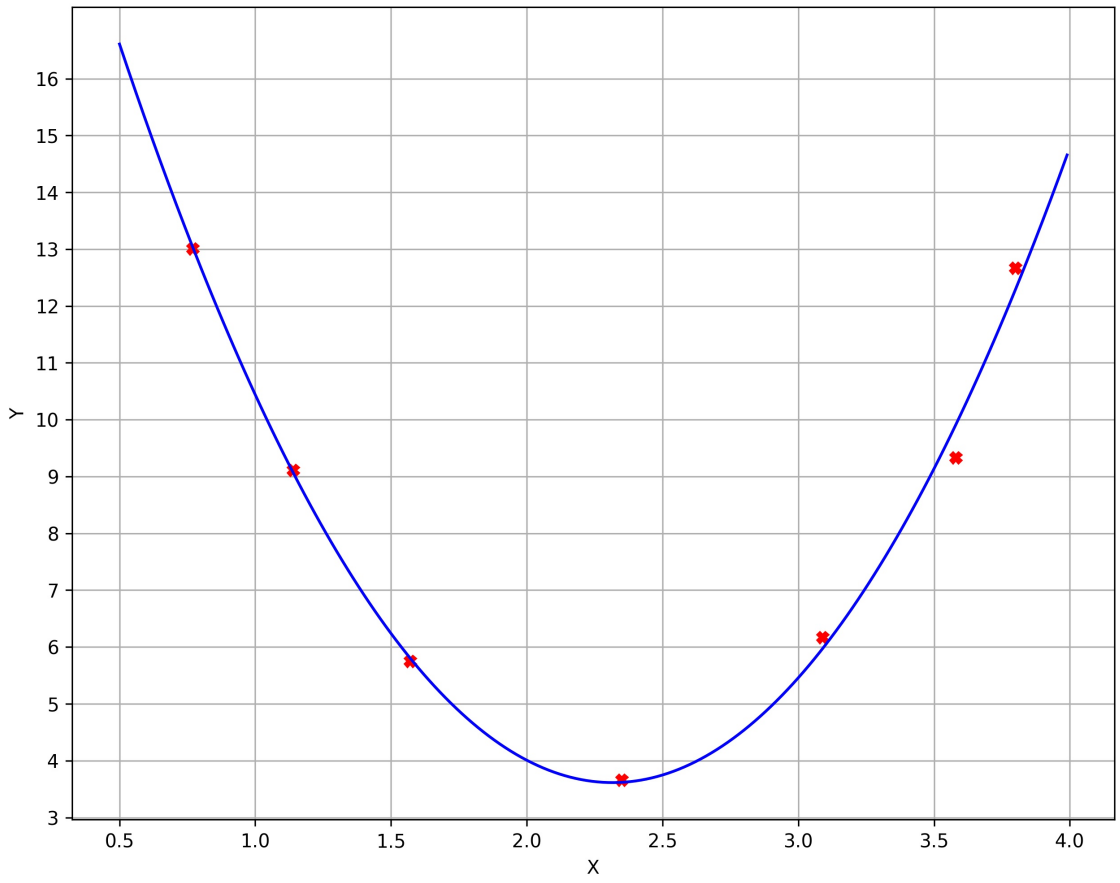
Знайдемо значення оцінок параметрів регресійної моделі за формулою:

$$\vec{\beta}^*_{\text{знач.}} = A^{-1}F^T\vec{\eta}_{\text{знач.}}$$

$\vec{\eta}_{\text{знач.}} = (13.01,\ \ 9.11,\ \ 5.75,\ \ 3.66,\ \ 6.17,\ \ 9.33,\ \ 12.67)^T$ Маємо: $\vec{\beta}^*_{\text{знач.}} = (24.74961129,\ \ -18.25186484,\ \ 3.94046229)^T$.
Знайшовши оцінки параметрів, отримуємо наступну модель:

$$f^*(x) = 24.74961129 - 18.25186484x + 3.94046229x^2$$

Зобразимо побудовану модель:



1.3 Перевірка адекватності побудованої моделі

Перевіримо побудовану модель на адекватність на рівні значущості $\alpha = 0.05$. Обчислимо значення виправленої вибіркової дисперсії $D^{**}\eta$ та залишкову оцінку дисперсії $(\sigma^2)^{**}$:

$$(D^{**}\eta)_{\text{знач.}} = \frac{1}{n-1}\sum_{i=1}^{n}(\eta_i - \bar{\eta})^2 \approx 12.533548$$

$$\left(\sigma^2\right)^{**}_{\text{знач.}} = \frac{1}{n-m}\left\|\vec{\eta} - F\vec{\beta}^*\right\|^2 \approx 0.13118$$

Обчислимо значення статистики F-критерію:

$$\zeta = \frac{D^{**}\eta}{(\sigma^2)^{**}} = \frac{\frac{1}{n-1}\sum_{k=1}^{n}(\eta_k - \bar{\eta})^2}{\frac{1}{n-m}\left\|\vec{\eta} - F\overrightarrow{\beta^*}\right\|} = \frac{\frac{1}{n-1}\sum_{k=1}^{n}(\eta_k - \bar{\eta})^2}{\frac{1}{n-m}\sum_{k=1}^{n}\left(\eta_k - f^*\left(\vec{x}^{(k)}\right)\right)^2}$$

$$\zeta_{\text{знач.}} = \frac{(D^{**}\eta)_{\text{знач.}}}{(\sigma^2)^{**}_{\text{знач.}}} = \frac{12.533548}{0.13118} \approx 95.5424$$

Знайдемо межу критичної області для $F(n-1, n-m)$, де n – кількість спостережень , а m – кількість невідомих параметрів.

У нашому випадку $n = 7, m = 3 \Rightarrow$ статистика ζ буде розподілена за законом Фішера-Снедекора із n - 1 = 6 та n - m = 4 ступенями вільності. Маємо $\zeta \sim F(6, 4)$

Критична область при саме такому виборі статистики критерію буде правостороння, тому знайдемо межу критичної області з виразу $\mathbb{P}(\zeta > t) = \alpha$. З таблиці розподілу Фішера-Снедекора знаходимо: $t_{\text{кр.}} = 6.16$. Маємо:

$$\zeta_{\text{знач.}} \approx 95.5424 > 6.16$$

Отже, $\zeta_{\text{знач.}} > t_{\text{кр.}}$ і модель можемо вважати адекватною на рівні значущості 0.05.

1.4 Перевірка гіпотези про значущість найменшого за значенням параметра

Перевіримо гіпотезу про значущість параметра для найменшого за значенням параметра $(\beta_2^*)_{\text{знач.}} = 3.94046229$ на рівні значущості $\alpha = 0.05$. Основна гіпотеза $H_0 : \beta_2 = 0$, альтернативна - $H_1 : \beta_2 > 0$, відповідно,

критична область - правостороння. Для перевірки гіпотези скористаємося статистикою:

$$\gamma = \frac{\beta_2^*}{\sqrt{(\sigma^2)^{**} a_4}} \sim St_{n-m}$$

З таблиці розподілу Стьюдента з $n - m = 4$ ступенями вільності отримуємо критичне значення $t_{\text{кр.}} = 2.132$. Значення компонентів формули:

$$(\beta_2^*)_{\text{знач.}} = 3.94046229$$
$$\left(\sigma^2\right)^{**}_{\text{знач.}} \approx 0.181096$$
$$a_{22} = \left(A^{-1}\right)_{22} \approx 4.5557084$$
$$\gamma_{\text{кр.}} \approx 23.8565 > t_{\text{кр.}}$$

Основна гіпотеза H_0 відхиляється на користь альтернативної. Таким чином, параметр β_2 є значущим, залишаємо побудовану модель незмінною.

1.5 Довірчий інтервал для середнього значення відклику та значення відклику в точці

Будуватимемо довірчий інтервал для середнього значення відклику та самого значення відклику в точці $x = 2$. Спочатку побудуємо прогнозований довірчий інтервал для середнього значення відклику з довірчою ймовірністю $\gamma = 0.95$. Для цього скористаємось статистикою :

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

Довірчий інтервал для середнього значення відклику матиме вигляд :

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

Величина t знаходиться з рівняння $\mathbb{P}(|\zeta| < t) = 0.95$; $t_{\text{зн.}} = 2.776$;
Таким чином, отримаємо довірчий інтервал для середнього значення відклику в точці x = 2 з надійністю $\gamma = 0.95$:

$$(-22.66241, 30.67787)$$

Тепер побудуємо довірчий інтервал для самого значення відклику в точці x = 2 з надійністю $\gamma = 0.95$. Для цього скористаємось статистикою:

$$\zeta = \frac{\eta - f^*(x)}{\sqrt{(\sigma^2)^{**}(1 + (\vec{x})^T A^{-1}\vec{x})}}$$

Довірчий інтервал матиме вигляд:

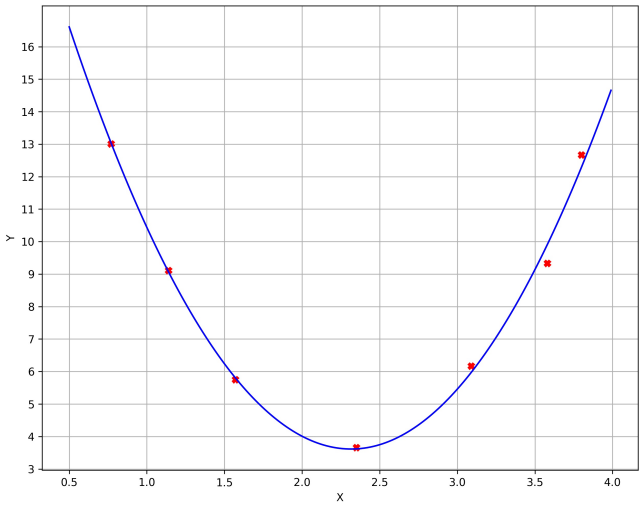
$$\left(f^*(x) - t\sqrt{(\sigma^2)^{**}(1 + (\vec{x})^T A^{-1}\vec{x})}, f^*(x) + t\sqrt{(\sigma^2)^{**}(1 + (\vec{x})^T A^{-1}\vec{x})}\right)$$

де t знаходимо з рівняння $\mathbb{P}(|\zeta| < t) = 0.95$. Всі необхідні нам значення були обчислені. Після усіх розрахунків отримаємо довірчий інтервал:

$$(-22.68135, 30.69681)$$

1.6 Висновки

В ході роботи було побудовано лінійну регресійну модель вигляду $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$. За допомогою методу найменших квадратів було знайдено оцінки параметрів цієї моделі. Була перевірена адекватність побудованої моделі на рівні значущості 0.05 та на тому ж самому рівні значущості була перевірена гіпотеза про значущість параметру з найменшим значенням його оцінки. Оскільки обидві гіпотези не протирічали дослідним даним, то модель залишилася незмінною. Був побудований довірчий інтервал для значення відклику та середнього значення відклику для значення фактору x = 2. Обидва інтервали були побудовані з довірчою ймовірністю 0.95. Обчислення були виконані за допомогою бібліотек Pandas, Matplotlib та Numpy мови програмування Python. Нижче на графіку зображена побудована лінія регресії:



2 Завдання 2

2.1 План роботи

1. За методом найменших квадратів знайти оцінки параметрів двофакторної регресійної моделі. Перевірити адекватність побудованої моделі.

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

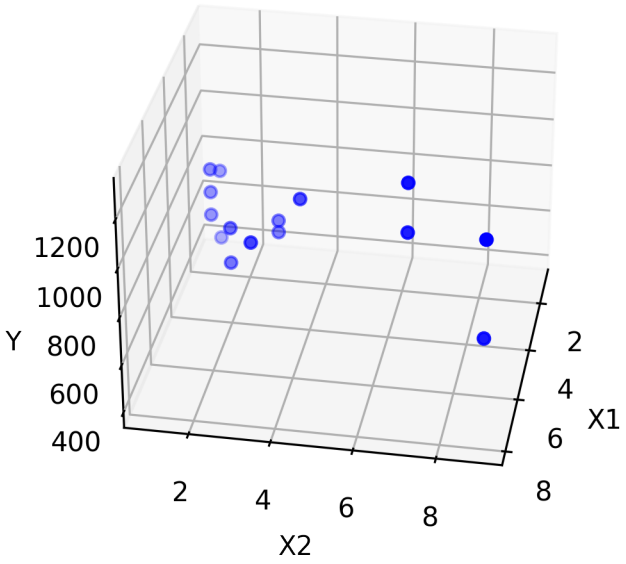
3. Для самого малого значення параметра побудованої моделі на рівні значущості $\alpha = 0.05$ перевірити гіпотезу про його значущість.

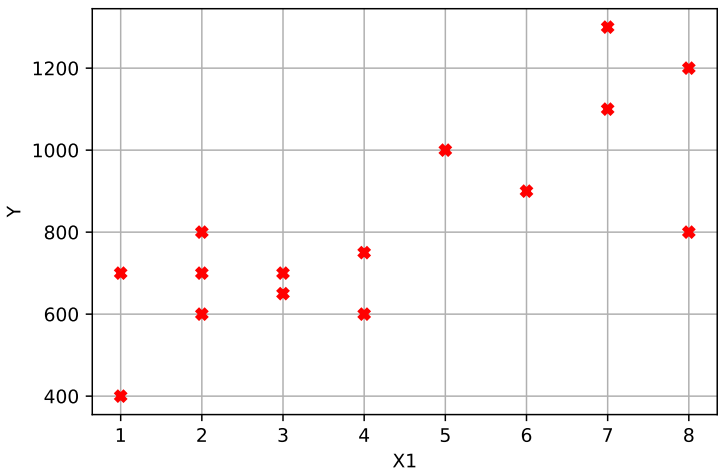
4. Побудувати прогнозований довірчий інтервал з довірчою ймовірністю $\gamma = 0.95$ для середнього значення відклику та самого значення відклику в деякій точці.(точку вибирайте самі).

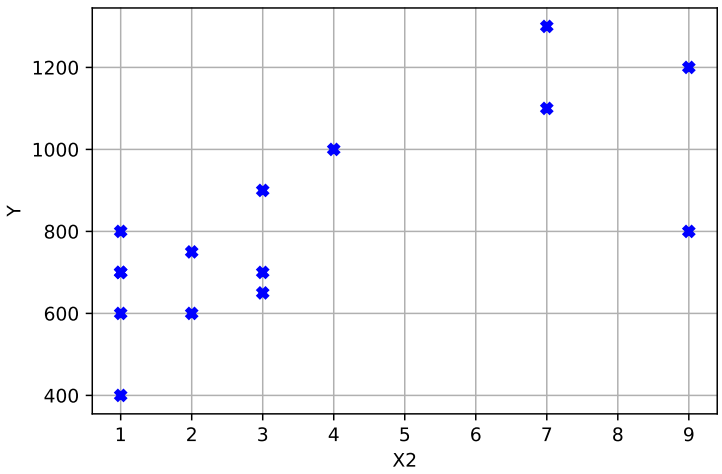
5. Написати висновки

Варіант № 13

№	X_1	X_2	Y
1	2	1	600
2	2	1	800
3	2	1	700
4	7	7	1300
5	7	7	1100
6	3	3	700
7	3	3	650
8	4	2	600
9	4	2	750
10	3	4	1000
11	8	9	800
12	8	9	1200
13	1	1	700
14	1	1	400
15	6	3	900


1200
1000
Y 800
600
400
2
4
6
8
X2
2
4
6
8
X1


1200
1000
800
600
400
Y
1
2
3
4
5
6
7
8
X1



Побудуємо просту двофакторну лінійну модель:
$f_1(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

2.2 Знаходження за МНК оцінки параметрів вибраної моделі

Матриця плану для вибраної моделі матиме вигляд:

$$F = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \cdots & 1 \\ 2 & 2 & 2 & 7 & 7 & 3 & 3 & 4 & \cdots & 6 \\ 1 & 1 & 1 & 7 & 7 & 3 & 3 & 2 & \cdots & 3 \end{pmatrix}^T$$

Після цього використовуємо матрицю плану F для знаходження інформаційної матриці Фішера: $A = FF^T$

$$A = \begin{pmatrix} 15 & 61 & 54 \\ 61 & 335 & 314 \\ 54 & 314 & 316 \end{pmatrix}^T$$

Знайдемо дисперсійну матрицю Фішера A^{-1}

$$A^{-1} = \begin{pmatrix} 0.29177 & -0.09319 & 0.04274 \\ -0.09319 & 0.07326 & -0.05688 \\ 0.04274 & -0.05688 & 0.05238 \end{pmatrix}^T$$

Знайдемо значення оцінок параметрів регресійної моделі за формулою:

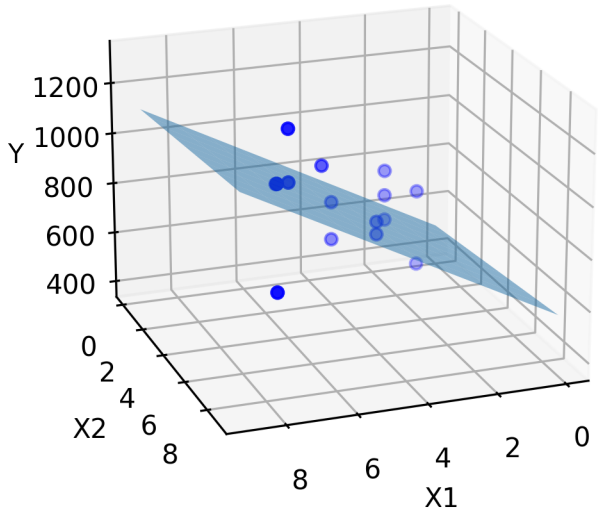
$$\vec{\beta}^*_{\text{знач.}} = A^{-1}F^T\vec{\eta}_{\text{знач.}},$$

$\vec{\eta}_{\text{знач.}} = (600,\ 800,\ 700,\ 1300,\ 1100, \cdots, 900)^T$

Отримаємо $\vec{\beta}^*_{\text{знач.}} \approx (544.64974293,\ 35.97365039,\ 33.99742931)^T$. Тоді побудована модель матиме вигляд:

$$f_1^*(x) = 544.64974293 + 35.97365039x_1 + 33.99742931x_2$$

Зобразимо графік отриманої моделі:



2.3 Перевірка адекватності побудованої моделі

Перевіримо побудовану модель на адекватність на рівні значущості $\alpha = 0.05$. Обчислимо значення виправленої вибіркової дисперсії $D^{**}\eta$ та залишкову оцінку дисперсії $(\sigma^2)^{**}$:

$$(D^{**}\eta)_{\text{знач.}} = \frac{1}{n-1}\sum_{i=1}^{n}(\eta_i - \bar{\eta})^2 \approx 60166.6667$$

$$\left(\sigma^2\right)^{**}_{\text{знач.}} = \frac{1}{n-m}\left\|\vec{\eta} - F\vec{\beta}^*\right\|^2 \approx 29864.9716$$

Обчислимо значення статистики F-критерію:

$$\zeta = \frac{D^{**}\eta}{(\sigma^2)^{**}} = \frac{\frac{1}{n-1}\sum_{k=1}^{n}(\eta_k - \bar{\eta})^2}{\frac{1}{n-m}\left\|\vec{\eta} - F\overrightarrow{\beta^*}\right\|} = \frac{\frac{1}{n-1}\sum_{k=1}^{n}(\eta_k - \bar{\eta})^2}{\frac{1}{n-m}\sum_{k=1}^{n}\left(\eta_k - f^*\left(\vec{x}^{(k)}\right)\right)^2}$$

$$\zeta_{\text{знач.}} = \frac{(D^{**}\eta)_{\text{знач.}}}{(\sigma^2)^{**}_{\text{знач.}}} = \frac{60166.6667}{29864.9716} \approx 2.01462$$

Знайдемо межу критичної області для $F(n-1, n-m)$, де n – кількість спостережень , а m – кількість невідомих параметрів.

У нашому випадку $n = 15, m = 3 \Rightarrow$ статистика γ буде розподілена за законом Фішера-Снедекора із 14 та 12 ступенями вільності. Маємо $\zeta \sim F(14, 12)$

Критична область при саме такому виборі статистики критерію буде правостороння, тому знайдемо межу критичної області з виразу $\mathbb{P}(\zeta > t) = \alpha$. З таблиці розподілу Фішера-Снедекора знаходимо: $t_{\text{кр.}} = 2.63712$. Маємо:

$$\zeta_{\text{знач.}} \approx 2.01462 < 2.63712$$

Бачимо, що $\zeta_{\text{знач.}} < t_{\text{кр.}}$, тому модель **не можна** вважати адекватною на рівні значущості 0.05.

3 Аналіз вибірки та побудова кращої регресійної моделі

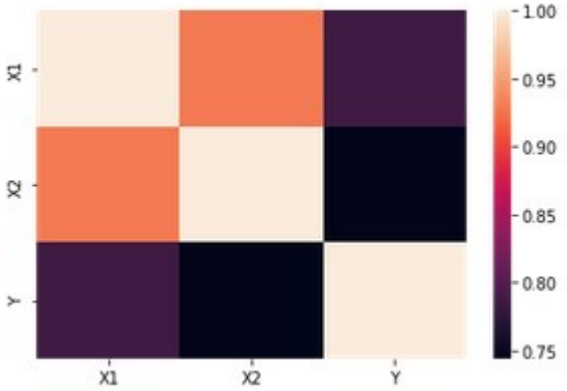
Щоб проаналізувати лінійну залежність між значеннями факторів та відклику, побудуємо нормовану кореляційну матрицю:

Для цього, використаємо формулу:

$\frac{\sum_{i=1}^{n}(\mathrm{x_i}-\overline{\mathrm{x}})(\mathrm{y_i}-\overline{\mathrm{y}})}{\sqrt{\sum_{i=1}^{n}(\mathrm{x_i}-\overline{\mathrm{x}})^2\sum_{i=1}^{n}(\mathrm{y_i}-\overline{\mathrm{y}})^2}}$, де $\overline{\mathrm{x}},\overline{\mathrm{y}}$ - середні значення,а x_i, y_i - значення факторів

$$R = \begin{pmatrix} 1 & 0.928781 & 0.786545 \\ 0.928781 & 1 & 0.744023 \\ 0.786545 & 0.744023 & 1 \end{pmatrix}^T$$

Візуалізуємо результати, які дозволять зробити висновки про лінійну залежність факторів та відклику. Коефіцієнт кореляції для значень фактора x_1 та відклику становить приблизно 0.786545, а для фактора x_2 та відклику - відповідно близко 0.744023



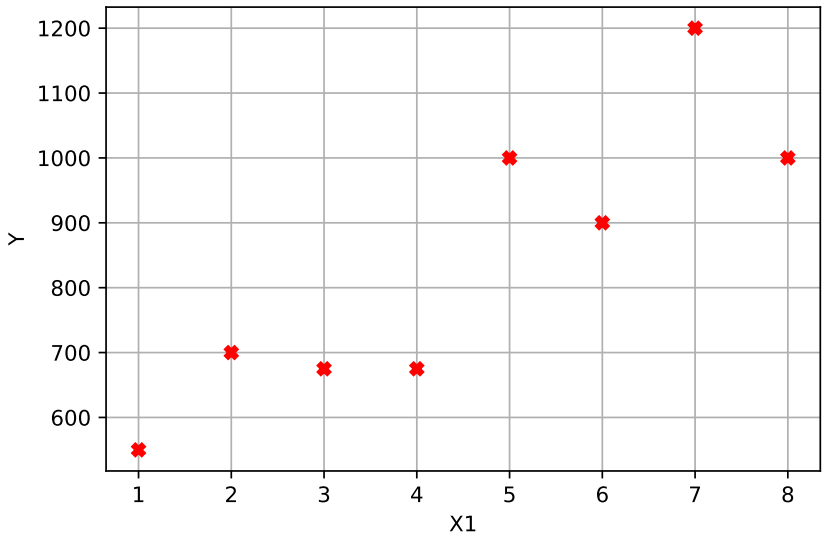
Звернемо увагу на значення факторів і відклику: бачимо, що присутні "дублікати"спостережень, при яких значення відклику різні. Це наштовхує на думку, що при проведенні вимірювань могла виникнути похибка. Для того, аби отримати більш точні значення, обчислимо середнє арифметичне значень відклику для однакових значень факторів. **Закон великих чисел** обґрунтовує такий підхід: нехай було зроблено декілька вимірів невипадкової величини Y та отримано випадкові значення $\eta_i, i \in \mathbb{N}$, близькі до Y. Будемо вважати, що в середньому отримані значення відкликів правильні (відсутність систематичних похибок). Тоді:

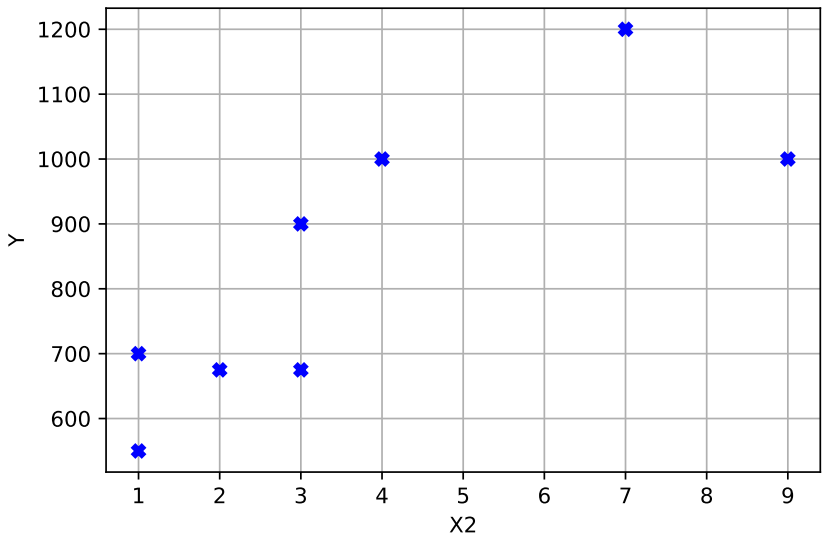
$$\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\eta_i = Y$$

$$\frac{1}{n}\sum_{k=1}^{n}\eta_i \xrightarrow{\text{P}} Y, n \to \infty$$

Отримуємо нові дослідні дані:

№	X_1	X_2	Y
1	2	1	700
2	7	7	1200
3	3	3	675
4	4	2	675
5	3	4	1000
6	8	9	1000
7	1	1	550
8	6	3	900

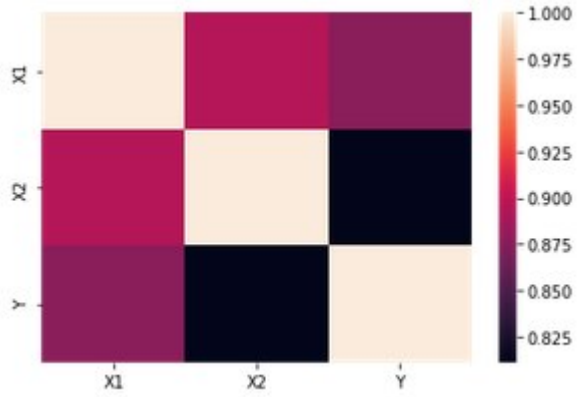



1200
1100
1000
900
800
700
600
Y
1
2
3
4
5
6
7
8
9
X2

Побудуємо нормовану кореляційну матрицю для нових дослідних даних.

$$R = \begin{pmatrix} 1 & 0.895353 & 0.876671 \\ 0.895353 & 1 & 0.811219 \\ 0.876671 & 0.811219 & 1 \end{pmatrix}^T$$

Отриманий коефіцієнт кореляції для значень фактора x_1 та відклику становить близько 0.876671, а для фактора x_2 та відклика - відповідно близько 0.811219.



Розглянемо модель:
$f_2(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

3.1 Знаходження за МНК оцінки параметрів вибраної моделі

Матриця плану для вибраної моделі матиме вигляд:

$$F = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 7 & 3 & 4 & 3 & 8 & 1 & 6 \\ 1 & 7 & 3 & 2 & 4 & 9 & 1 & 3 \end{pmatrix}^T$$

Після цього використовуємо матрицю плану F для знаходження інформаційної матриці Фішера: $A = FF^T$

$$A = \begin{pmatrix} 8 & 34 & 30 \\ 34 & 188 & 171 \\ 30 & 171 & 170 \end{pmatrix}^T$$

Знайдемо дисперсійну матрицю Фішера A^{-1}

$$A^{-1} = \begin{pmatrix} 0.55809 & -0.13342 & 0.03571 \\ -0.13342 & 0.09442 & -0.07143 \\ 0.03571 & -0.07143 & 0.07143 \end{pmatrix}^T$$

Знайдемо значення оцінок параметрів регресійної моделі за формулою:

$$\vec{\beta}^*_{\text{знач.}} = A^{-1}F^T\vec{\eta}_{\text{знач.}},$$

$\vec{\eta}_{\text{знач.}} = (700,\ 1200,\ 675,\ 675,\ 1000,\ 1000,\ 550,\ 900)^T$ Отримаємо $\vec{\beta}^*_{\text{знач.}} \approx (565.82512315,\ 26.10837438,\ 42.85714286)^T$. Знайшовши оцінки параметрів, отримуємо модель:

$$f_2^*(x) = 565.82512315 + 26.10837438x_1 + 42.85714286x_2 \qquad (1)$$

3.2 Перевірка адекватності побудованої моделі

Перевіримо побудовану модель на адекватність на рівні значущості $\alpha = 0.05$. Обчислимо значення виправленої вибіркової дисперсії $D^{**}\eta$ та залишкову оцінку дисперсії $(\sigma^2)^{**}$:

$$(D^{**}\eta)_{\text{знач.}} = \frac{1}{n-1}\sum_{i=1}^{n}(\eta_i - \bar{\eta})^2 \approx 45870.5357$$

$$\left(\sigma^2\right)^{**}_{\text{знач.}} = \frac{1}{n-m}\left\|\vec{\eta} - F\vec{\beta}^*\right\|^2 \approx 9989.9239$$

Обчислимо значення статистики F-критерію:

$$\zeta = \frac{D^{**}\eta}{(\sigma^2)^{**}} = \frac{\frac{1}{n-1}\sum_{k=1}^{n}(\eta_k - \bar{\eta})^2}{\frac{1}{n-m}\left\|\vec{\eta} - F\overrightarrow{\beta^*}\right\|} = \frac{\frac{1}{n-1}\sum_{k=1}^{n}(\eta_k - \bar{\eta})^2}{\frac{1}{n-m}\sum_{k=1}^{n}\left(\eta_k - f^*\left(\vec{x}^{(k)}\right)\right)^2}$$

$$\zeta_{\text{знач.}} = \frac{(D^{**}\eta)_{\text{знач.}}}{(\sigma^2)^{**}_{\text{знач.}}} = \frac{45870.5357}{9989.9239} \approx 4.59168$$

Знайдемо межу критичної області для $F(n-1, n-m)$, де n – кількість спостережень , а m – кількість невідомих параметрів.

У нашому випадку $n = 8, m = 3 \Rightarrow$ статистика γ буде розподілена за законом Фішера-Снедекора із 7 та 5 ступенями вільності. Маємо $\zeta \sim F(7, 5)$

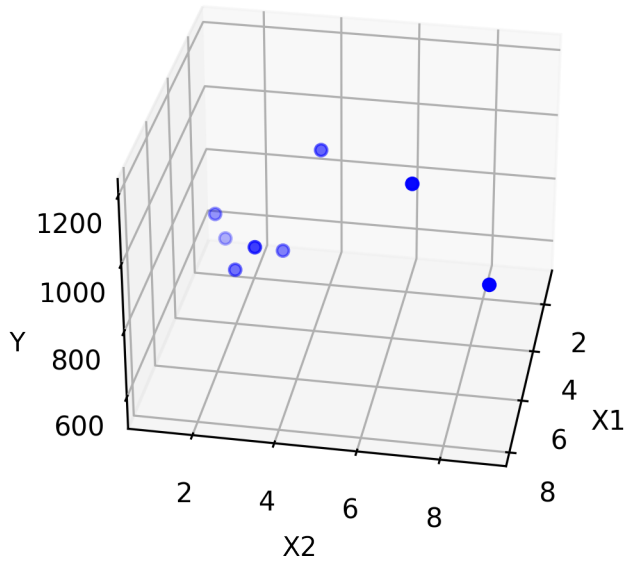
Критична область при саме такому виборі статистики критерію буде правостороння, тому знайдемо межу критичної області з виразу $\mathbb{P}(\zeta > t) = \alpha$. З таблиці розподілу Фішера-Снедекора знаходимо: $t_{\text{кр.}} = 4.8759$ Маємо:

$$\zeta_{\text{знач.}} \approx 4.59168 < 4.8759$$

Таким чином, $\zeta_{\text{знач.}} < t_{\text{кр.}}$, тому модель **не можна** вважати адекватною на рівні значущості 0.05.

4 Аналіз вибірки та побудова кращої регресійної моделі

З огляду на розташування точок-значень двох факторів та відклику у тривімірному просторі, зробимо припущення, що залежність може бути квадратичною:



Розглянемо модель:
$f_3(x) = \beta_0 + \beta_2 x_2 + \beta_1 x_1^2 + \beta_3 x_1 x_2$

4.1 Знаходження за МНК оцінки параметрів вибраної моделі

Матриця плану для вибраної моделі матиме вигляд:

$$F = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 7 & 3 & 2 & 4 & 9 & 1 & 3 \\ 4 & 49 & 9 & 16 & 9 & 64 & 1 & 36 \\ 2 & 49 & 9 & 8 & 12 & 72 & 1 & 18 \end{pmatrix}^T$$

Після цього використовуємо матрицю плану F для знаходження інформаційної матриці Фішера: $A = FF^T$

$$A = \begin{pmatrix} 8 & 30 & 188 & 171 \\ 30 & 170 & 1127 & 1139 \\ 188 & 1127 & 8228 & 7983 \\ 171 & 1139 & 7983 & 8203 \end{pmatrix}^T$$

Знайдемо дисперсійну матрицю Фішера A^{-1}

$$A^{-1} = \begin{pmatrix} 1.23717 & -0.60019 & -0.03392 & 0.09056 \\ -0.60019 & 0.38127 & 0.01282 & -0.0529 \\ -0.03392 & 0.01282 & 0.00326 & -0.00424 \\ 0.09056 & -0.0529 & -0.00424 & 0.00971 \end{pmatrix}^T$$

Знайдемо значення оцінок параметрів регресійної моделі за формулою:

$$\vec{\beta}^*_{\text{знач.}} = A^{-1}F^T\vec{\eta}_{\text{знач.}},$$

$\vec{\eta}_{\text{знач.}} = (700,\ 1200,\ 675,\ 675,\ 1000,\ 1000,\ 550,\ 900)^T$ Отримаємо $\vec{\beta}^*_{\text{знач.}} \approx (407.2399665,\ 172.92047169,\ 10.07709773,\ -21.28671494)^T$. Тоді сама модель матиме вигляд:

$$f^*_3(x) = 407.2399665 + 172.92047169x_2 + 10.07709773x_1^2 - 21.28671494x_1x_2$$

4.2 Перевірка адекватності побудованої моделі

Перевіримо побудовану модель на адекватність на рівні значущості $\alpha = 0.05$. Обчислимо значення виправленої вибіркової дисперсії $D^{**}\eta$ та залишкову оцінку дисперсії $(\sigma^2)^{**}$:

$$(D^{**}\eta)_{\text{знач.}} = \frac{1}{n-1}\sum_{i=1}^{n}(\eta_i - \bar{\eta})^2 \approx 45870.5357$$

$$\left(\sigma^2\right)^{**}_{\text{знач.}} = \frac{1}{n-m}\left\|\vec{\eta} - F\vec{\beta}^*\right\|^2 \approx 6363.1592$$

Обчислимо значення статистики F-критерію:

$$\zeta = \frac{D^{**}\eta}{(\sigma^2)^{**}} = \frac{\frac{1}{n-1}\sum_{k=1}^{n}(\eta_k - \bar{\eta})^2}{\frac{1}{n-m}\left\|\vec{\eta} - F\overrightarrow{\beta^*}\right\|} = \frac{\frac{1}{n-1}\sum_{k=1}^{n}(\eta_k - \bar{\eta})^2}{\frac{1}{n-m}\sum_{k=1}^{n}\left(\eta_k - f^*\left(\vec{x}^{(k)}\right)\right)^2}$$

$$\zeta_{\text{знач.}} = \frac{(D^{**}\eta)_{\text{знач.}}}{(\sigma^2)^{**}_{\text{знач.}}} = \frac{45870.5357}{6363.1592} \approx 7.20877$$

Знайдемо межу критичної області для $F(n-1, n-m)$, де n – кількість спостережень , а m – кількість невідомих параметрів.

У нашому випадку $n = 8, m = 4 \Rightarrow$ статистика γ буде розподілена за законом Фішера-Снедекора із 7 та 4 ступенями вільності. Маємо $\zeta \sim F(7, 4)$

Критична область при саме такому виборі статистики критерію буде правостороння, тому знайдемо межу критичної області з виразу $\mathbb{P}(\zeta > t) = \alpha$. З таблиці розподілу Фішера-Снедекора знаходимо: $t_{\text{кр.}} = 6.0942$ Маємо:

$$\zeta_{\text{знач.}} \approx 7.20877 > 6.0942$$

Бачимо, що $\zeta_{\text{знач.}} > t_{\text{кр.}}$, тому модель **можна вважати адекватною** на рівні значущості 0.05.

4.3 Перевірка гіпотези про значущість найменшого за значенням параметра

Перевіримо гіпотезу про значущість параметра для $(\beta_2^*)_{\text{знач.}} = 10.07709773$ на рівні значущості $\alpha = 0.05$. Основна гіпотеза $H_0 : \beta_2 = 0$, альтернативна $H_1 : \beta_2 > 0$, відповідно, критична область - правостороння. Для перевірки гіпотези скористаємося статистикою:

$$\gamma = \frac{\beta_2^*}{\sqrt{(\sigma^2)^{**} a_{22}}} \sim St_4$$

З таблиці розподілу Стьюдента з $n - m = 4$ ступенями вільності отримуємо критичне значення $t_{\text{кр.}} = 2.132$. Знаходимо $(\sigma^2)^{**} \approx 6363.1592$, $a_{22} = (A^{-1})_{22} \approx 0.00326$. Отримуємо $\gamma_{\text{знач.}} \approx 2.21411 > 2.132$
Оскільки $\gamma_{\text{знач.}} > t_{\text{кр.}}$, то основна гіпотеза H_0 відхиляється на користь альтернативної. Таким чином, параметр β_2 є значущим, залишаємо побудовану модель незмінною.

4.4 Довірчий інтервал для середнього значення відклику та значення відклику в точці

Будуватимемо довірчий інтервал для середнього значення відклику та самого значення відклику в точці (1, 1).
Спочатку побудуємо прогнозований довірчий інтервал для середнього значення відклику з довірчою ймовірністю $\gamma = 0.95$. Для цього скористаємось статистикою :

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

Довірчий інтервал для середнього значення відклику матиме вигляд :

$$\left(f^*(\vec{x}) - t\sqrt{(\sigma^2)^{**}(\vec{x})^T A^{-1}\vec{x}}, f^*(\vec{x}) + t\sqrt{(\sigma^2)^{**}(\vec{x})^T A^{-1}\vec{x}}\right) \tag{3}$$

Величина t знаходиться з рівняння $\mathbb{P}(|\zeta| < t) = 0.95;\ t_{\text{зн.}} = 2.776;$
Таким чином, отримаємо довірчий інтервал для середнього значення відклику в точці (1, 1) з надійністю $\gamma = 0.95$:

$$(419.47495, 718.42669) \tag{4}$$

Тепер побудуємо довірчий інтервал для самого значення відклику в точці (1, 1) з надійністю $\gamma = 0.95$. Для цього скористаємось статистикою:

$$\zeta = \frac{\eta - f^*(\vec{x})}{\sqrt{(\sigma^2)^{**}(1 + (\vec{x})^T A^{-1}\vec{x})}} \sim St_{n-m} = St_4 \tag{5}$$

Довірчий інтервал матиме вигляд:

$$\left(f^*(\vec{x}) - t\sqrt{(\sigma^2)^{**}(1 + (\vec{x})^T A^{-1}\vec{x})}, f^*(\vec{x}) + t\sqrt{(\sigma^2)^{**}(1 + (\vec{x})^T A^{-1}\vec{x})}\right) \tag{6}$$

де t знаходимо з рівняння $\mathbb{P}(|\zeta| < t) = 0.95$. Всі необхідні нам значення були обчислені. Після усіх розрахунків отримаємо довірчий інтервал:

$$(301.78297, 836.11867)$$

4.5 Висновки

В ході роботи було побудовано лінійну регресійну модель вигляду: $f_1(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. За допомогою методу найменших квадратів було знайдено оцінки параметрів моделі. Було з'ясовано, що дана модель не є адекватною на рівні значущості 0.05, тобто не є кращою за константну. Оскільки у вибірці містилося багато повторень "дублікатів"у знаначеннях факторів із різними значеннями відклику, було вирішено взяти середні значення відклику та обґрунтувати даний підхід з допомогою закону великих чисел (ЗВЧ). Побудовано нову двофакторну просту лінійну модель вигляду $f_2(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ За допомогою методу найменших квадратів було знайдено оцінки параметрів моделі. Було з'ясовано, що дана модель також не є адекватною на рівні значущості 0.05. Зроблено припущення, що залежність між значеннями факторів та відклику є квадратичною. Побудовано модель $f_3(x) = \beta_0 + \beta_2 x_2 + \beta_1 x_1^2 + \beta_3 x_1 x_2$. З'ясовано що таку модель можна вважати адекватною на рівні значущості 0.05. На тому ж самому рівні значущості була перевірена гіпотеза про значущість параметру з найменшим значенням його оцінки. Оскільки гіпотеза не протирічили дослідним даним, то дана модель залишилася незмінною. В подальшому був побудований довірчий інтервал для значення відклику та середнього значення відклику в точці (1, 1). Обидва інтервали були побудовані з довірчою ймовірністю 0.95.

Примітка 1. Для початкових дослідних даних була побудована модель: $f_4(x) = \beta_0 + \beta_2 x_2 + \beta_1 x_1^2 + \beta_3 x_1 x_2$ Було з'ясовано, що її не можна вважати адекватною на рівні значущості 0.05, оскільки:

$$\zeta_{\text{знач.}} \approx 2.05398 < 2.63712 = t_{\text{кр.}}$$

Отже, похибка вимірювань, яка могла бути зроблена, значним чином впливає на значення оцінок параметрів для даної моделі. Варто зазначити, що мала кількість спостережень у вибірці може призвести до виникнення "перенавчання ситуації, коли модель «дуже добре вивчає» (в термінах Machine Learning) специфіку заданих на початку даних, але немає гарантії, що отримані в майбутньому спостереження матимуть таку саму особливість. Таким чином, збільшення степеня поліноміальної моделі або зменшення кількості спостережень на практиці може бути неправильним підходом.

Примітка 2 Можна зробити висновки, що обрана модель має недоліки через невідому природу наданої вибірки. Можна припустити наявність інших, не наданих важливих факторів, з урахуванням яких дані моделі будуть гірше працювати для передбачення нових данних, в такій ситуації побудована модель може бути неефективною (виникнення згаданого вище "перенавчання").

Список використаної літератури

[1] "Конспект лекцій з теорії ймовірностей та математичної статистики"– Каніовська І.Ю.
[2] "Математическая статистика"– издательство МГТУ им. Н.Э. Баумана

Список використаного програмного забезпечення

1. TeXstudio (LaTeX Editor)
2. Jupyter Notebook
3. Python Libraries: SciPy, NumPy, Matplotlib