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

РОЗРАХУНКОВА РОБОТА

з теми "Регресійний аналіз"

Виконав:
студент групи КА-91
Панаско Віталій

Перевірила:
Каніовська І.Ю.

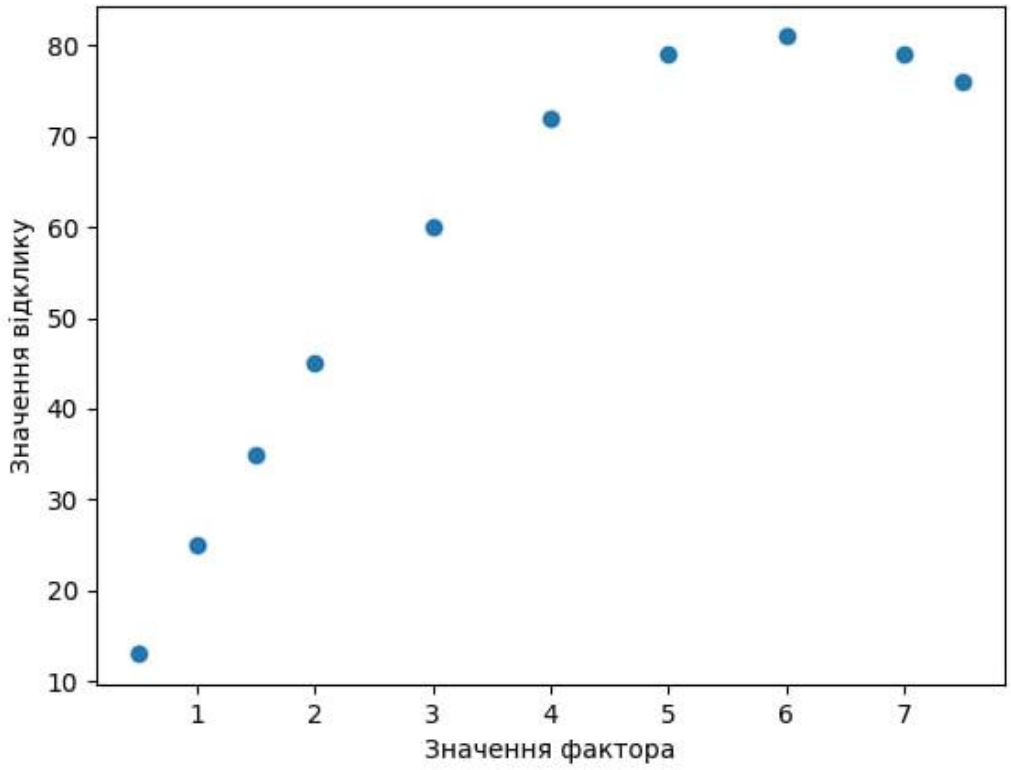
Київ 2021

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 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.
6. Написати висновки.

x_i	0,5	1	1,5	2	3	4	5	6	7	7,5
y_i	13	25	35	45	60	72	79	81	79	76

1.1 Провести аналіз вибірки та вибрати підходящу лінійну регресійну модель



За розташуванням точок – пар значень (фактор-відклик) бачимо, що вони розміщені на площині не лінійно, а більше нагадують параболу. Тому доцільно буде взяти модель вигляду

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

Внаслідок такого вибору виникає ряд запитань. Чому ми беремо максимальний саме степінь 2, а не більший? Зокрема можемо взяти степінь 3, 4 або набагато більший, наприклад 20. Інше запитання в тому, чи можемо ми вибрати інші базисні функції та отримати більш точну модель. Зокрема ми можемо взяти базисну пару $\{1, Asin(Bx + C)\}$ чи $\{1, D\sqrt{x + E}\}$, де A,B,C,D,E в кожному випадку оцінюємо попередньо. Відповідаю по порядку.

- Базові функції. Тут все доволі просто: «поліноміальна» лінійна модель вважається стандартною лінійною моделлю, а самі поліноми – добре вивчені дослідниками. Крім того, така модель є доволі простою та

інтуїтивно зрозумілою. Інші базисні функції використовуються в дещо складніших задачах, зокрема тригонометричні функції (синус, косинус) можуть використовуватись в моделях з певною «періодичною» особливістю.

- Степінь. Насправді тут не існує чіткого правила, яку степінь вибирати для «поліноміальної» лінійної моделі. Існують лише деякі рекомендації. Збільшення степеня призводить до більшої «точності» лінії регресії на наборі початкових даних. Іншими словами пряма проходить «практично»(дуже близько) через кожну точку на діаграмі розсіювання, що дані нам по умові задачі. Але такий підхід є неправильним, адже така модель досліджує і «вивчає» (в термінах Machine Learning) специфіку саме початкових, заданих даних. Це називається **переобучение**(рос.) моделі. Нема ніякої гарантії, що дані, які ми отримаємо в ході наступних досліджень матимуть таку особливість, тому в якості прогнозу для значень відклику таку складну модель не розглядають. Так само, доречі, як і занадто просту модель, але тут ситуація навпаки – модель є апріорі занадто простою, а отже й неточною. Це називається **недообучение**(рос.) моделі.

  Взагалі кажучи, нам треба обрати компроміс між цими варіантами[2] Тобто обрати модель середню по «складності» (complexibillity англ.) Це призведе до наступних результатів (рис. 1):

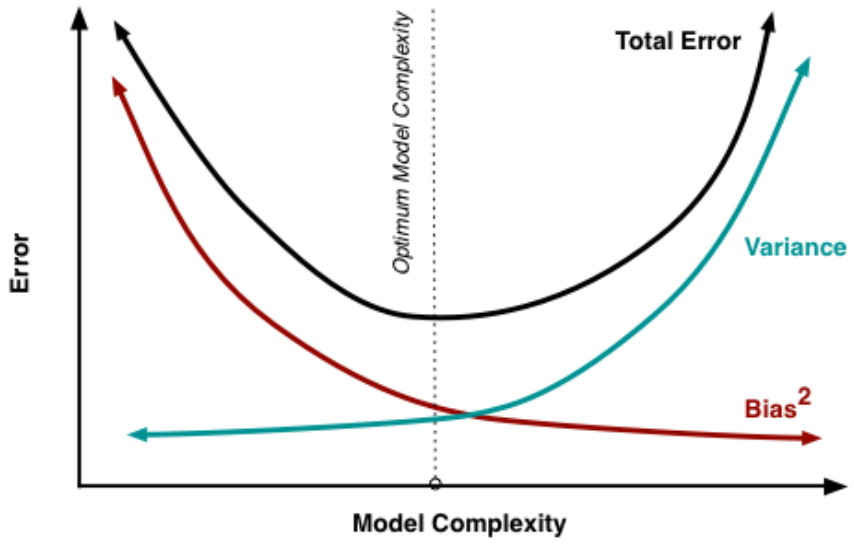


Рис.1

Дана проблема вибору називається **Bias-Variance Tradeoff**[1]. Фактично, обравши такий шлях вибору, ми будемо знаходитись на низькому рівні **Bias**(різниця від реального значення відклику та прогнозованого) та **Variance**(дисперсії від умовного мат. сподівання, тобто $Df(x)$, що

позначає розсіювання прогнозованих значень відносно реальних). Проілюструємо ці результати(рис. 2):

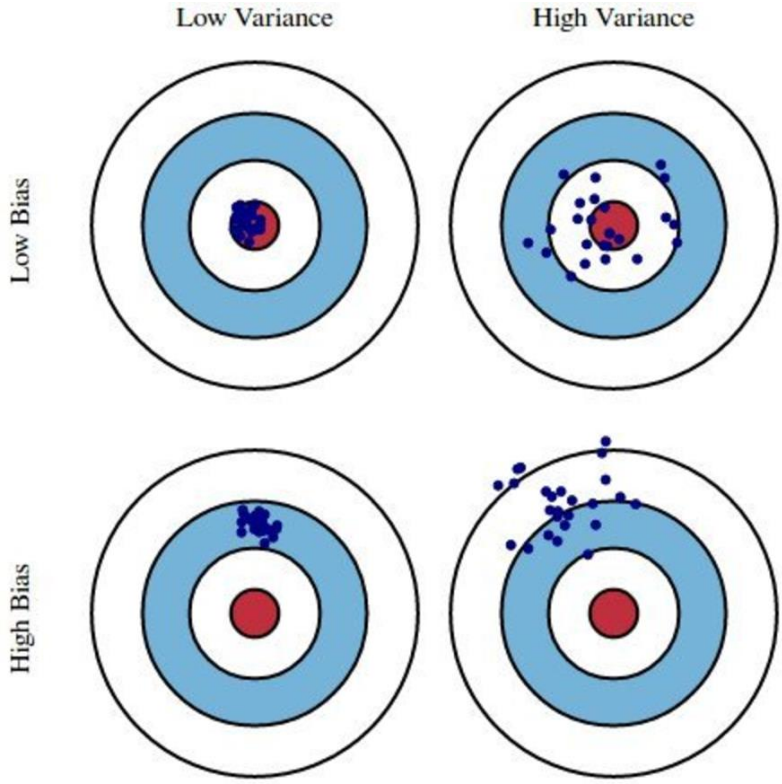


Рис. 2

Використовуючи попередні терміни, **переобучение** – призведе до низького Bias на даних(початкових та скоріш за все й на прогнозованих), але великого Variance(розсіювання) прогнозованих даних відносно своїх реальних значень. Тоді як **недообучение** дасть нам невелике розсіювання, але велике відхилення від реальних значень.

В контексті нашого завдання стандартна лінійна модель вигляду $f(x) = \beta_0 + \beta_1 x$ призведе до **недообучения**, а більш складні «поліноміальні» з великим максимальним ступенем до **переобучения**. Я пропоную піти на компроміс та розлянути саме «квадратичну» лінійну модель, яка є «середньою» за складністю моделлю та простою в подальших обчисленнях.

1.2 Знаходження за МНК оцінки параметрів вибраної моделі

Матриця плану для вибраної моделі матиме вигляд:

$$F = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 0.5 & 1 & \ldots & 7.5 \\ 0.5^2 & 1^2 & \cdots & 7.5^2 \end{pmatrix}^T$$

В загальному випадку, щоб знайти оцінку за допомогою МНК варто зробити припущення, що $rangF = 3$ та $\vec{\varepsilon} \sim N(\vec{0}, \sigma^2 I)$. В нашому випадку припущення про ранг справджується, отже робимо припущення лише про розподіл похибок.

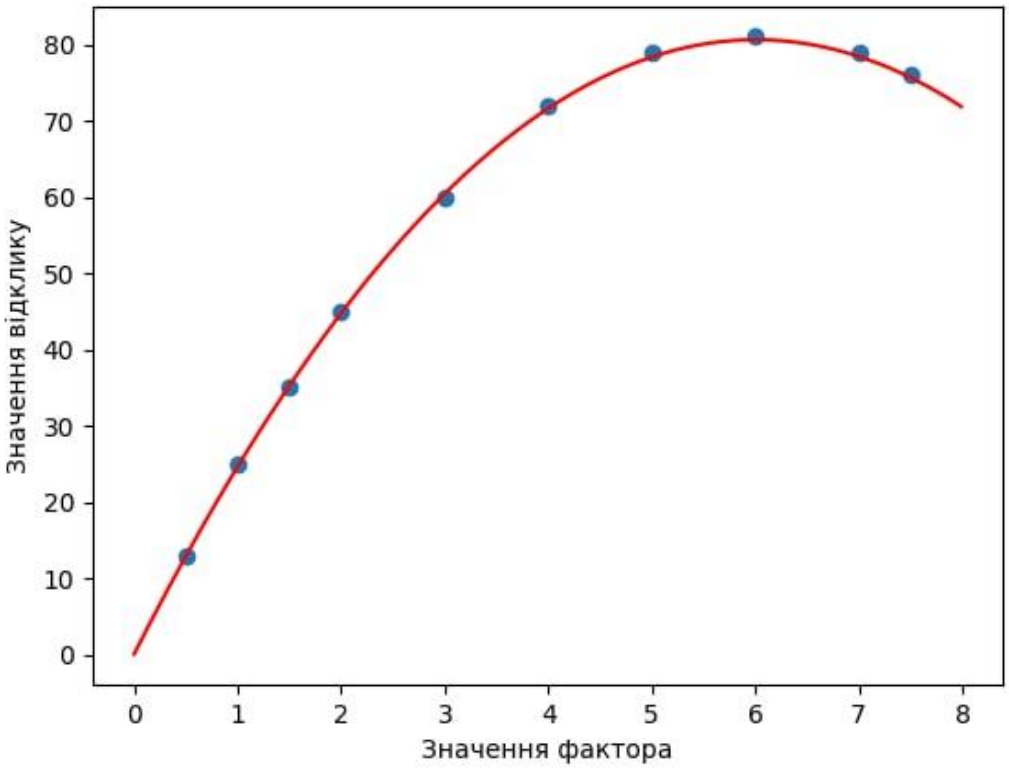
Далі переходимо до обрахунків: використовуємо матрицю F для знаходження інформаційної матриці Фішера $A = F^T F$ та дисперсійну матрицю Фішера A^{-1}; обидві з них мають розмір 3×3.

$$A = \begin{pmatrix} 10 & 37.5 & 199 \\ 37.5 & 199 & 1210 \\ 199 & 1210 & 7850 \end{pmatrix}; A^{-1} = \begin{pmatrix} 0.875 & -0.487 & 0.053 \\ -0.487 & 0.353 & -0.042 \\ 0.053 & -0.042 & 0.00526 \end{pmatrix};$$

$$\overrightarrow{\eta_{\text{зн}}} = (13, 25, \ldots, 76); \overrightarrow{\beta^*_{\text{зн}}} = A^{-1} F^T \overrightarrow{\eta_{\text{зн}}} \approx \begin{pmatrix} 0.0858 \\ 26.8 \\ -2.23 \end{pmatrix}$$

Знайшовши оцінки параметрів, отримуємо наступну модель:

$$f^*(x) = 0.0858 + 26.8x - 2.23x^2$$



Зображення побудованої моделі на діаграмі розсіювання

1.3 Перевірка адекватності побудованої моделі на рівні значущості $\alpha = 0.05$

Ми вже знаємо, що

$$\gamma = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(\eta_i - \bar{\eta})^2}{\frac{1}{n-m}\left\|\vec{\eta} - F\vec{\beta}\right\|^2} \sim F(n-1, n-m)$$

Висуваємо наступні гіпотези:

H_0: *константа модель та побудована не відрізняються*

H_1: *побудована модель є кращою за константну*

Для нашого випадку $n = 10, m = 3 \Longrightarrow \gamma \sim F(9, 7)$

Не забуваємо, що область правостороння; з таблиць знаходимо:

$$t_{\text{кр}} = 3.68$$

Обчислимо значення $D^{**}\eta$ та $(\sigma^2)^{**}$:

$$D^{**}\eta_{\text{зн}} \approx 636$$
$$(\sigma^2)^{**}{}_{\text{зн}} \approx 0.234$$

$$\gamma_{\text{зн}} \approx 2718 > t_{\text{кр}}$$

Ми потрапили в критичну область, отже на рівні значущості $\alpha = 0.05$ дані протирічать гіпотезі про «однаковість» моделей; H_0 відхиляємо, H_1 приймаємо і модель вважаємо адекватною на цьому ж рівні значущості.

1.4 Перевірка гіпотези про значущість найменшого за значенням параметра

Будемо перевіряти гіпотезу про значущість найменшого за значенням параметра β_0 на тому ж рівні значущості $\alpha = 0.05$

$$\beta_0{}^{*}{}_{\text{зн}} = 0.0858$$

Висуваємо наступні гіпотези:

$$H_0: \beta_0 = 0$$
$$H_1: \beta_0 > 0$$

Очевидно, критична область правостороння. В якості статистики критерію використаємо:

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

Для нашого випадку $n = 10, m = 3 \Longrightarrow \gamma \sim St_7$

З таблиць знаходимо: $t_{\text{кр}} = 1.895$

Значення компонентів формули:

$$\beta_0{}^*{}_{\text{зн}} = 0.0858$$
$$(\sigma^2)^{**}{}_{\text{зн}} \approx 0.234$$
$$a_{00} = (A^{-1})_{00} \approx 0.875$$

$$\gamma_{\text{зн}} \approx 0.1896 < t_{\text{кр}}$$

Ми потрапили в область прийняття гіпотези, отже припущення про те, що параметр β_0 є незначущим на рівні значущості $\alpha = 0.05$ ми приймаємо. Таким чином модифікуємо нашу модель:

$$f^*(x) = 26.8x - 2.23x^2$$

Перевіримо її на адекватність:

$$n = 10, m = 2 \Longrightarrow \gamma \sim F(9, 8)$$

З таблиць знаходимо:

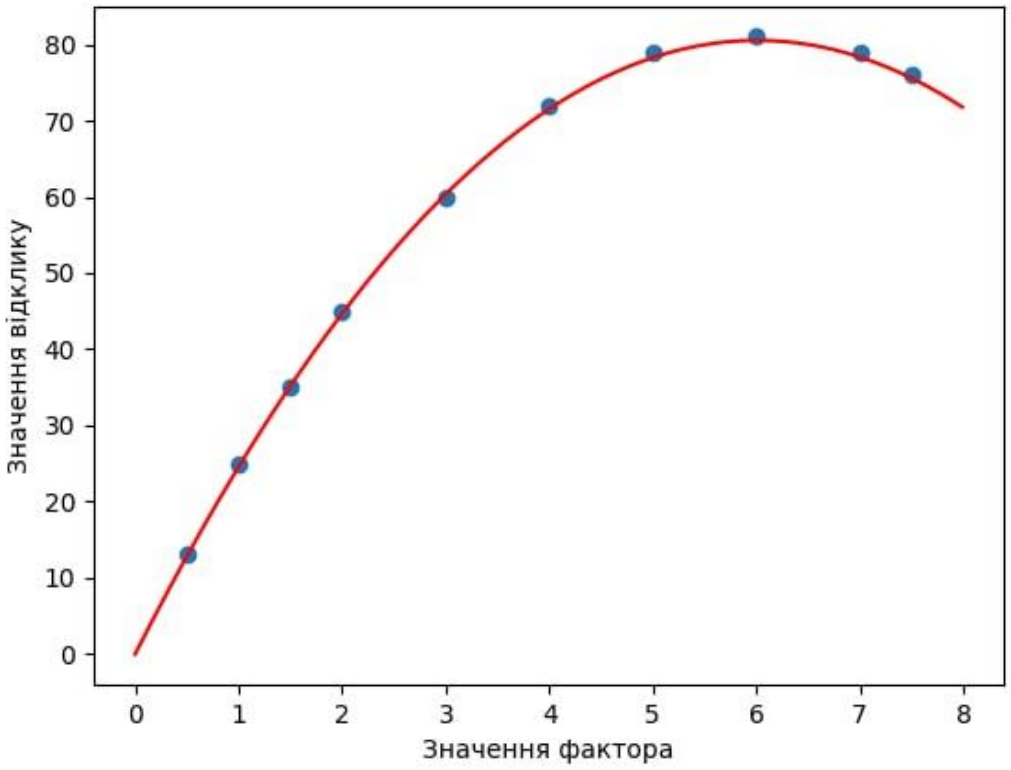
$$t_{\text{кр}} = 3.39$$

Змінилось лише значення $(\sigma^2)^{**}$:

$$(\sigma^2)^{**}{}_{\text{зн}} \approx 0.30256$$

$$\gamma_{\text{зн}} \approx 2102 > t_{\text{кр}}$$

Ми потрапили в критичну область, отже на рівні значущості $\alpha = 0.05$ дані не протирічать гіпотезі про адекватність модифікованої моделі (аналогічно до попереднього пункту).



Зображення модифікованої моделі на діаграмі розсіювання

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

Спочатку визначимося з точкою для якої будемо будувати довірчі інтервали: $x_0 = 6.5$ з довірчою ймовірністю $\gamma = 0.95$. $\vec{x}$ тут вважатимемо вже вибраним набором значень факторів(без втрати загальності)

Далі обчислимо значення $(\overrightarrow{x})^T A^{-1}\vec{x}$, яке знанобиться нам в обох формулах:

$$(\overrightarrow{x})^T A^{-1}\vec{x} = (1, 6.5, 6.5^2)\begin{pmatrix} 0.875 & -0.487 & 0.053 \\ -0.487 & 0.353 & -0.042 \\ 0.053 & -0.042 & 0.00526 \end{pmatrix}\begin{pmatrix} 1 \\ 6.5 \\ 6.5^2 \end{pmatrix} \approx 0.25767$$

- Інтервал для середнього значення

$$\frac{f^*(\vec{x}) - f(\vec{x})}{\sqrt{(\sigma^2)^{**}(\overrightarrow{x})^T A^{-1}\vec{x}}} \sim St_8$$

Шуканий довірчий інтервал має вигляд:

$$f(\vec{x}) \in \left(f^*(\vec{x}) - t\sqrt{(\sigma^2)^{**}_{\text{зн}}(\overrightarrow{x})^T A^{-1}\vec{x}}; f^*(\vec{x}) + t\sqrt{(\sigma^2)^{**}_{\text{зн}}(\overrightarrow{x})^T A^{-1}\vec{x}}\right)$$

Де t знаходимо з таблиці квантилів розподілу Стьюдента:

$$t = 2.306$$

Компоненти формули:

$$f^*(\vec{x}) = 79.9825$$

$$(\sigma^2)^{**}_{\text{зн}} \approx 0.234$$

$$(\overrightarrow{x})^T A^{-1}\vec{x} \approx 0.25767$$

Остаточно маємо:

$$\left(79.9825 - 2.306\sqrt{0.234 \cdot 0.25767}; 79.9825 + 2.306\sqrt{0.234 \cdot 0.25767}\right) \approx (79.41626, 80.548738)$$

- Інтервал для значення відклику

$$\frac{\eta - f^*(\vec{x})}{\sqrt{(\sigma^2)^{**}(1 + (\overrightarrow{x})^T A^{-1}\overrightarrow{x})}} \sim St_8$$

Шуканий довірчий інтервал має вигляд:

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

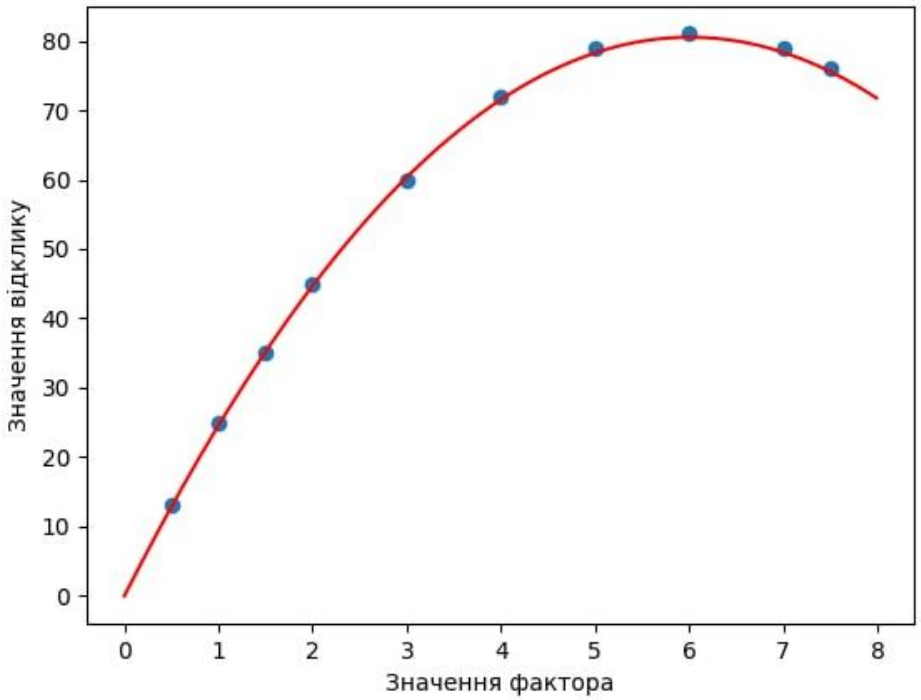
Компоненти формули ми вже знаємо, залишилося порахувати:

$$\left(79.9825 - 2.306\sqrt{0.234 \cdot (1 + 0.25767)}; 79.9825 + 2.306\sqrt{0.234 \cdot (1 + 0.25767)}\right) \approx (78.731519, 81.23348)$$

1.6 Висновки

В ході роботи над завданням 1 було побувано квадратичну лінійну модель вигляду $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$. Вибір саме такої лінійної моделі був зроблений для забезпечення середньої "складності" (англ. complexibility)

моделі з огляду на проблему Bias-Variance Tradeoff[1]: саме така по своїй складності модель забезпечить низьке відхилення від очікуваного значення відклику і реального(low Bias) та низьку дисперсію для $f(\vec{x})$(слабке розсіювання очікуваного значення відклику відносно реального) (low Variance). За допомогою методу найменших квадратів було знайдено оцінки для параметрів моделі. Було перевірено гіпотезу про адекватність побудованої моделі та гіпотезу про значущість найменшого за значенням параметра(в нашому випадку для β_0) на заданному рівні значущості. Якщо в першому випадку дані не протирічили гіпотезі про адекватність, то в другому виявилось, що параметр β_0 виявився незначущим. Після модифікації побудованої моделі, ми знову перевірили її на адекватність та отримали позитивний результат. Також було побудовано довірчі інтервали для значення відклику та середнього значення відклику у вибраній точці з довірчої ймовірністю $\gamma = 0.95$. Внизу зображено діаграму розсіювання разом з побудованою модифікованою моделлю.

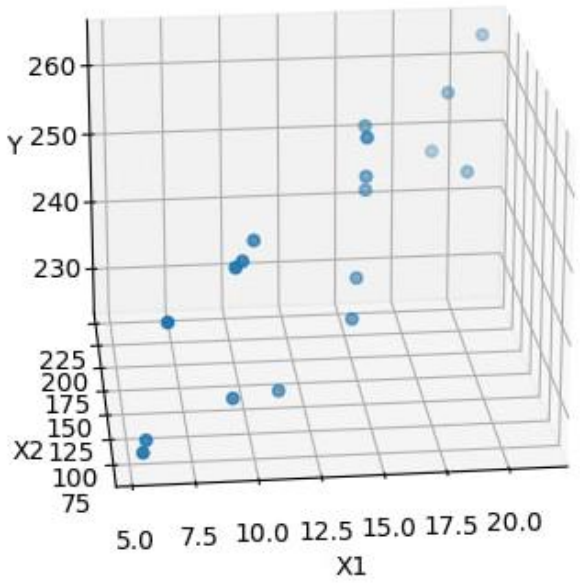


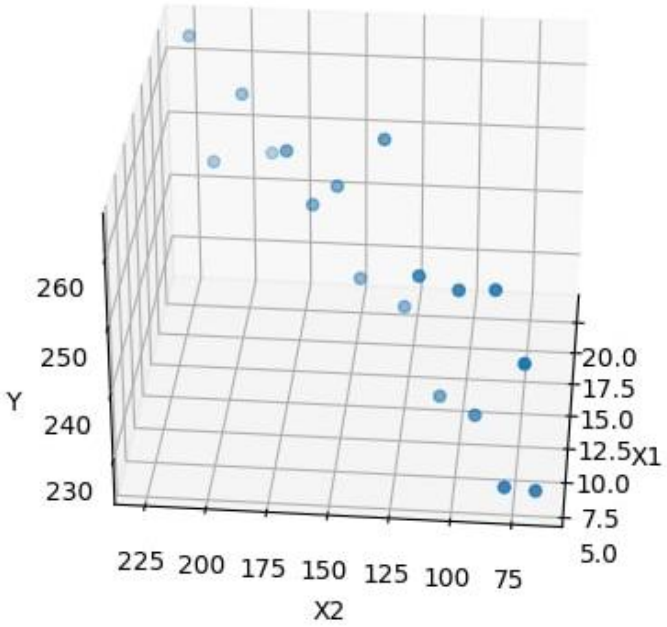
Завдання 2

1.За методом найменших квадратів знайти оцінки параметрів двофакторної регресійної моделі.
2.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.
6. Написати висновки

№	X_1	X_2	Y
1	5.50	66.00	226.16
2	6.60	72.20	243.11
3	5.70	78.80	226.05
4	9.40	85.80	248.45
5	9.30	93.20	229.38
6	9.80	101.00	247.38
7	11.30	109.20	227.96
8	10.40	117.80	248.10
9	14.50	126.80	235.37
10	15.20	136.20	259.58
11	14.90	146.00	238.71
12	15.40	156.20	251.85
13	15.50	166.80	248.67
14	15.60	177.80	256.53
15	20.10	189.20	248.12
16	19.40	201.00	258.27
17	18.90	213.20	248.50
18	21.20	225.80	263.80

На малюнку нижче зображено кілька ракурсів тривимірної діаграми розсіювання.





З огляду на те, що точки рівномірно збільшують свою висоту по мірі збільшення значень одночасно по осям X1 та X2, побудуємо просту двофакторну лінійну модель

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

2.1 Знаходження за МНК оцінки параметрів вибраної моделі

Матриця плану для вибраної моделі матиме вигляд:

$$F = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 5.5 & 6.6 & \ldots & 21.2 \\ 66 & 72.2 & \cdots & 225.8 \end{pmatrix}^T$$

В загальному випадку, щоб знайти оцінку за допомогою МНК варто зробити припущення, що $rangF = 3$ та $\vec{\varepsilon} \sim N(\vec{0}, \sigma^2 I)$. В нашому випадку припущення про ранг знову, як і в завданні 1, справджується, отже робимо припущення лише про розподіл похибок.

Далі переходимо до обрахунків: використовуємо матрицю F для знаходження інформаційної матриці Фішера $A = F^T F$ та дисперсійну матрицю Фішера A^{-1}; обидві з них мають розмір 3 × 3.

$$A = \begin{pmatrix} 18 & 238.7 & 2463 \\ 238.7 & 3584.13 & 36778.4 \\ 2463 & 36778.4 & 380244.36 \end{pmatrix};$$

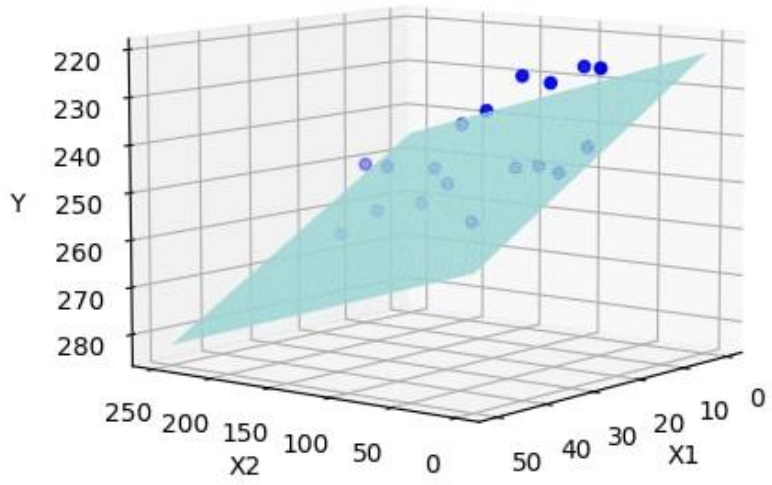
$$A^{-1} \approx \begin{pmatrix} 0.49071 & -0.00862 & -0.00234 \\ -0.00862 & 0.03745 & -0.00357 \\ -0.00234 & -0.00357 & 0.0003628 \end{pmatrix};$$

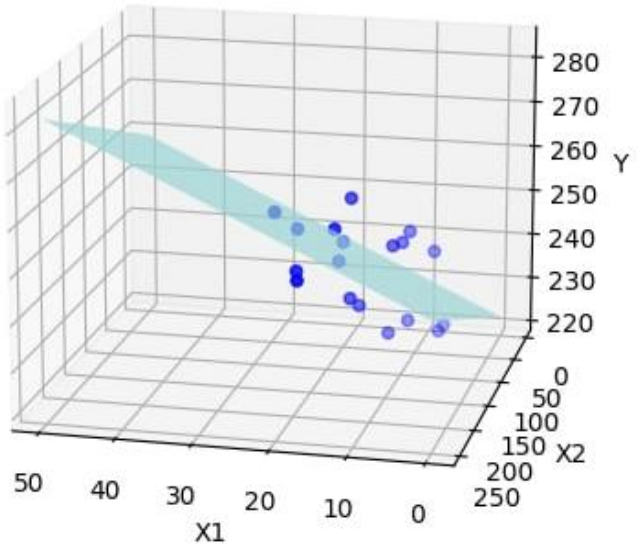
$$\overrightarrow{\eta_{\text{зн}}} = (226.16, 243.11, \ldots, 263.80);$$

$$\overrightarrow{\beta^*_{\text{зн}}} = A^{-1} F^T \overrightarrow{\eta_{\text{зн}}} \approx \begin{pmatrix} 222.17 \\ 0.73453 \\ 0.094031 \end{pmatrix}$$

Знайшовши оцінки параметрів, отримуємо наступну модель:

$$f^*(\vec{x}) = 222.17 + 0.73453x_1 + 0.094031x_2$$





2.2 Перевірка адекватності побудованої моделі на рівні значущості $\alpha = 0.05$

Ми вже знаємо, що

$$\gamma = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(\eta_i - \bar{\eta})^2}{\frac{1}{n-m}\left\|\vec{\eta} - F\vec{\beta}\right\|^2} \sim F(n-1, n-m)$$

Висуваємо наступні гіпотези:

H_0: *константа модель та побудована не відрізняються*

H_1: *побудована модель є кращою за константну*

Для нашого випадку $n = 18, m = 3 \Longrightarrow \gamma \sim F(17, 15)$

Не забуваємо, що область правостороння; з Excel знаходимо:

$$t_{\text{кр}} = 2.36827$$

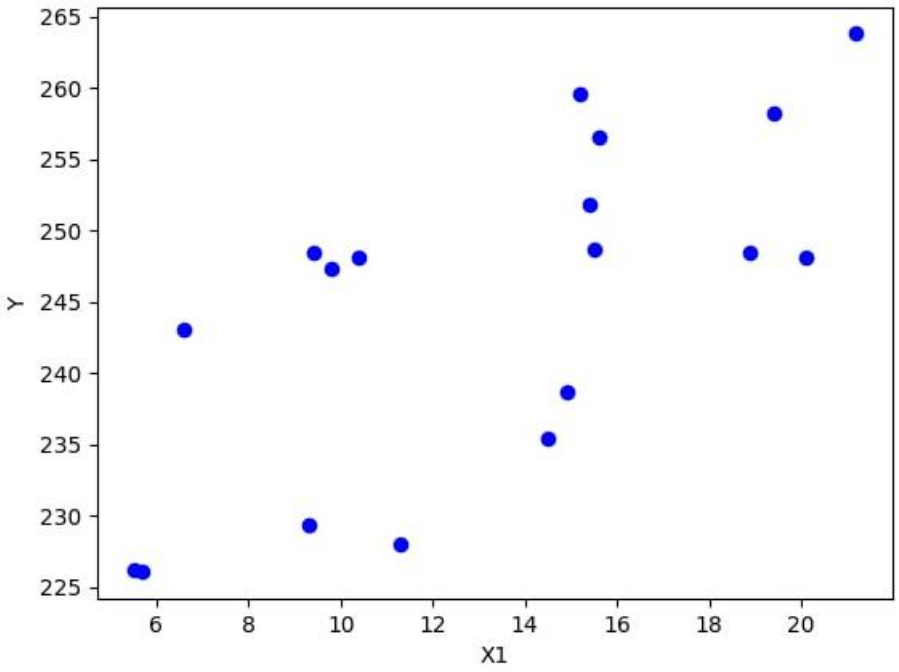
Обчислимо значення $D^{**}\eta$ та $(\sigma^2)^{**}$:

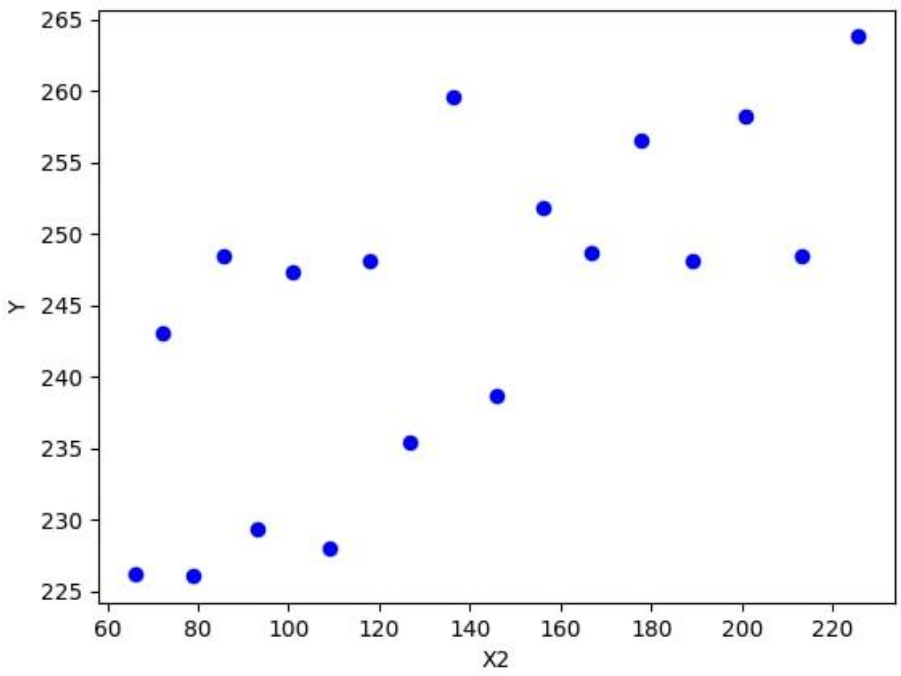
$$D^{**}\eta_{\text{зн}} \approx 139.29275$$
$$(\sigma^2)^{**}{}_{\text{зн}} \approx 79.419$$

$$\gamma_{\text{зн}} \approx 1.75389 < t_{\text{кр}}$$

Ми отримали, що на рівні значущості $\alpha = 0.05$ дані протирічать гіпотезі про адекватність моделі. В такому випадку існує два варіанти виходу: перший – це підібрати іншу модель, зокрема підбираючи інші фактори (вибрати інші базисні функції для факторів замість лінійних), другий дуже простий – підвищити рівень значущості. Я пропоную піти саме другим шляхом і зараз поясню чому.

Насправді я не думаю, що нам вдасться підібрати кращу модель: якщо проста лінійна регресія між x_1 та y простежується доволі складно і там можна побачити навіть константну регресійну модель, то у випадку з x_2 та y вона майже очевидна; можна зробити припущення, що так як лінії регресії "для площин" x_1oy та x_2oy є прямими, то сама модель буде мати вигляд площини, що перерізає ці площини по лініям регресії. А це якраз той випадок, який ми вже розглянули та навіть знайшли "найкращі" коефіцієнти – для площини.





Які ж можуть бути "модифікації" площини? Я уже згадав, що між x_1 та y трохи простежується проста константна "залежність": точки розсіяні доволі хаотично. Теоретично ми можемо знехтувати β_1 та побудувати площину регресії як просту нахилену під деяким кутом до $x_1 o x_2$ площину. На практиці я отримав гірші результати, зокрема я використовував ті ж знайдені коефіцінти і вважав, що $\beta_1 = 0$; значення статистики $(\sigma^2)^{**}$ вийшло ще більшим, а отже ми тим паче не проходимо F-тест. Інший крок – це побудувати просту однофакторну лінійну модель для x_2 та y, повністю знехтувавши x_1 та знайти нові коефіцієнти; але це невиправданий крок, адже ми достеменно не знаємо "вплив" x_1 на y.

Ще один можливий варіант – проста константа модель. На користь цього варіанту, до речі, говорить той факт, що, незважаючи на зміну значень x_1 та x_2 від 0 до 25 та від 0 до 250 відповідно, значення y змінилось лише від 225 до 255, тобто фактично лише на 30 одиниць. Але це не вихід тим паче, адже $\gamma_{зн}$ буде = 1 і F-тест ми знову не проходимо. Не даремно рекомендують додавати фактори, а не спрощувати модель. З приводу більш "складних" моделей лише скажу, що, аналізуючи діаграму розсіювання, висота все-таки збільшується рівномірно по мірі одночасного зростання значень x_1 і x_2 та ніякої складної фігури чи поверхні (наприклад параболоїд чи ін.) я там не побачив. Тому я пропоную не ускладнювати модель додатковими факторами та обчисленнями та прийняти побудовану площину як адекватну модель, але з вищим рівнем значущості. Тим паче значення, $t_{кр} = 2.36827$ не дуже сильно відрізняється від $\gamma_{зн} \approx 1.75389$ і це можна пояснити зокрема незначною похибкою отримання даних.

$$t_{\text{кр},0.15} = 1.714014 < \gamma_{\text{зн}} = 1.75389$$

Висновок: на рівні значущості $\alpha = 0.15$ дані не суперечать гіпотезі про адекватність побудованої моделі.

2.3 Перевірка гіпотези про значущість найменшого за значенням параметра

Будемо перевіряти гіпотезу про значущість найменшого за значенням параметра β_2 на тому ж рівні значущості, на якому довели адекватність моделі $\alpha = 0.15$

$$\beta_2{}^*{}_{\text{зн}} = 0.094031$$

Висуваємо наступні гіпотези:

$$H_0: \beta_2 = 0$$
$$H_1: \beta_2 > 0$$

Критична область правостороння:

$$\gamma = \frac{\beta^*{}_2}{\sqrt{(\sigma^2)^{**} \cdot a_{22}}} \sim St_{15}$$

З таблиць знаходимо: $t_{\text{кр}} = 1.073531$

Значення компонентів формули:

$$\beta_2{}^*{}_{\text{зн}} = 0.094031$$
$$(\sigma^2)^{**}{}_{\text{зн}} \approx 79.419$$

$$a_{22} = (A^{-1})_{22} \approx 0.0003628$$

$$\gamma_{\text{зн}} \approx 0.553956 < t_{\text{кр}}$$

Ми потрапили в область прийняття гіпотези, отже припущення про те, що параметр β_0 є незначущим на рівні значущості $\alpha = 0.15$ ми приймаємо. Таким чином модифікуємо нашу модель:

$$f^*(\vec{x}) = 222.17 + 0.73453x_1$$

Перевіримо її на адекватність:

$$n = 18, m = 2 \Longrightarrow \gamma \sim F(17, 16)$$

З таблиць знаходимо:

$$t_{\text{кр},0.15} = 2.3167$$

Змінилось лише значення $(\sigma^2)^{**}$:

$$(\sigma^2)^{**}{}_{\text{зн}} \approx 284.5826$$

$$\gamma_{\text{зн}} \approx 0.489463 < t_{\text{кр}}$$

Висновок: на рівні значущості $\alpha = 0.15$ дані протирічать гіпотезі про адекватність модифікованої моделі і тому я пропоную повернутися до початково побудованої моделі. Тут ми могли знову підвищити рівень значущості, але на мою думку робити це не варто: в нас вже є адекватна модель на нижчому рівні значущості.

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

Спочатку визначимося з точкою для якої будемо будувати довірчі інтервали: $x_0 = (15, 160)$ з довірчою ймовірністю $\gamma = 0.95$. $\vec{x}$ тут вважатимемо вже вибраним набором значень факторів(без втрати загальності)

Далі обчислимо значення $(\overrightarrow{x})^T A^{-1}\vec{x}$, яке знанобиться нам в обох формулах:

$$(\overrightarrow{x})^T A^{-1}\vec{x} = (1, 15, 160)\begin{pmatrix} 0.49071 & -0.00862 & -0.00234 \\ -0.00862 & 0.03745 & -0.00357 \\ -0.00234 & -0.00357 & 0.0003628 \end{pmatrix}\begin{pmatrix} 1 \\ 15 \\ 160 \end{pmatrix} \approx 0.06124$$

- Інтервал для середнього значення

$$\frac{f^*(\vec{x}) - f(\vec{x})}{\sqrt{(\sigma^2)^{**}(\overrightarrow{x})^T A^{-1}\vec{x}}} \sim St_{15}$$

Шуканий довірчий інтервал має вигляд:

$$f(\vec{x}) \in \left(f^*(\vec{x}) - t\sqrt{(\sigma^2)^{**}{}_{\text{зн}}(\overrightarrow{x})^T A^{-1}\vec{x}}; f^*(\vec{x}) + t\sqrt{(\sigma^2)^{**}{}_{\text{зн}}(\overrightarrow{x})^T A^{-1}\vec{x}}\right)$$

Де t знаходимо з таблиці квантилів розподілу Стьюдента:

$$t = 2.131$$

Компоненти формули:

$$f^*(\vec{x}) = 248.23291$$

$$(\sigma^2)^{**}_{\text{зн}} \approx 79.419$$

$$(\overrightarrow{x})^T A^{-1}\vec{x} \approx 0.06124$$

Остаточно маємо:

$$\left(248.23291 - 2.131\sqrt{79.419 \cdot 0.06124}; 248.23291 + 2.131\sqrt{79.419 \cdot 0.06124}\right) \approx (243.53328\ 252.93253)$$

- Інтервал для значення відклику

$$\frac{\eta - f^*(\vec{x})}{\sqrt{(\sigma^2)^{**}(1 + (\overrightarrow{x})^T A^{-1}\overrightarrow{x})}} \sim St_8$$

Шуканий довірчий інтервал має вигляд:

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

Компоненти формули ми вже знаємо, залишилося порахувати:

$$\left(248.23291 - 2.131\sqrt{79.419 \cdot (1 + 0.06124)}; 248.23291 + 2.131\sqrt{79.419 \cdot (1 + 0.06124)}\right)$$

$$\approx (228.66914, 267.79667)$$

2.5 Висновки

В ході роботи над завданням 2 було побудовано просту лінійну двофакторну модель. Вибір саме такої лінійної моделі був зроблений в результаті аналізу трьохвимірної діаграми розсіювання: "висота" точок поступово і плавно зростала по мірі одночасного збільшення значень факторів. Крім того розміщення точок на діаграмі не "нагадувало" складну фігуру чи поверхню, а більше походило на звичайну площину. Далі було побудовано оцінки параметрів моделі за МНК. Після цього була перевірка

на адекватність моделі і виявилось, що на заданому рівні значущості $\alpha = 0.05$ гіпотезу про адекватність ми змушені відхилити. Я вирішив піти шляхом збільшення рівня значущості, зважаючи на недоречність змінювати (додавання факторів і.т.д.) інтуїтивно найкращу модель. На рівні значущості $\alpha = 0.15$ дані вже не суперечили висунутій гіпотезі, зважаючи на це я вирішив перевірити на цьому ж рівні гіпотезу про значущість найменшого за значенням параметра, а саме β_2. Дані не суперечили гіпотезі про незначущість параметра, тож я модифікував модель. Проте перевірка на адекватність новопобудованої моделі на цьому ж рівні дала негативний результат: я повернувся до початкової: "адекватність" нової вимагала б ще вищого рівня значущості. В самому кінці я побудував довірчі інтервали для середнього значення відклику та самого значення відклику у вибраній точці з довірчою ймовірністю $\gamma = 0.95$.

Використані джерела:

- https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff
- https://alexguanga.github.io/biasvariance.html
- https://www.machinelearningmastery.ru/linear-vs-polynomial-regression-walk-through-83ca4f2363a3/
- https://www.machinelearningmastery.ru/polynomial-regression-bbe8b9d97491/
- Електронний конспект лекцій – Каніовська І.Ю.
- Python; бібліотеки NumPy, Matplotlib, Pandas