## Приклад дослідження рівняння парної лінійної регресії

**Приклад 6.1.** Для аналізу залежності об'єма споживання $Y$ (у.о.) домогосподарств від доходу $X$ (у.о.) відібрана вибірка об'єму $n=12$ (помісячно протягом року), результати якої наведені в таблиці.

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_i$ | 107 | 109 | 110 | 113 | 120 | 122 | 123 | 128 | 136 | 140 | 145 | 150 |
| $y_i$ | 102 | 105 | 108 | 110 | 115 | 117 | 119 | 125 | 132 | 130 | 141 | 144 |

Вивчається залежність між випадковими величинами $X$ і $Y$.

Необхідно:

1) побудувати діаграму розсіювання;

2) визначити тип залежності;

3) оцінити параметри рівняння регресії $Y$ на $X$;

4) нанести на діаграму розсіювання графік функції регресії;

5) оцінити силу лінійної залежності між $X$ і $Y$;

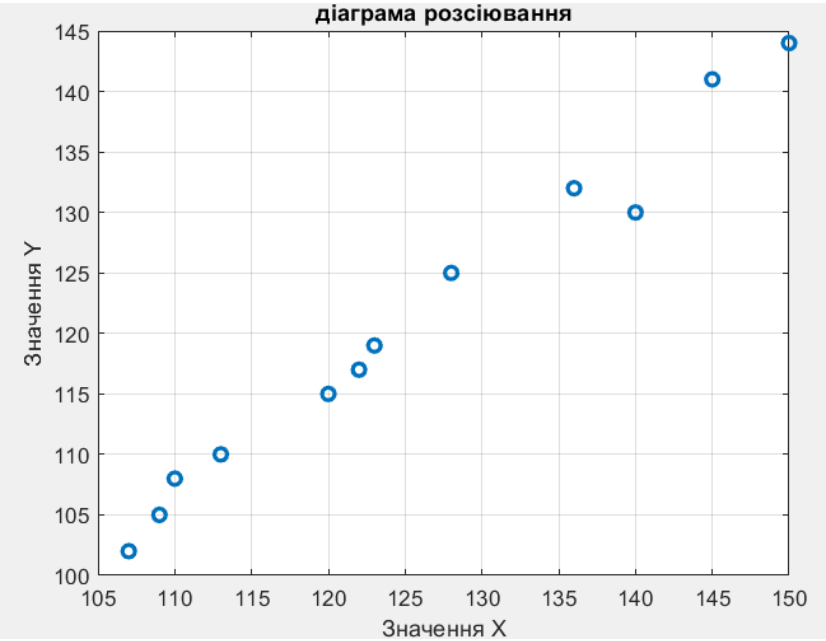
6) оцінити статистичну значущість коефіцієнтів регресії при рівні значущості $\alpha=0,05$;

7) розрахувати 95%-ті довірчі інтервали для теоретичних коефіцієнтів регресії;

8) спрогнозувати результативну ознаку $X=x_p$ ($x_p=160$); визначити 95%-ті довірчі інтервали для цього прогнозу.

Розв'язання:

1) побудувати діаграму розсіювання;




2) визначити тип залежності;

За розташуванням точок на кореляційному полі припускаємо, що залежність між $X$ та $Y$ лінійна: $y = b_0 + b_1 \cdot x$.

3) оцінити параметри рівняння регресії $Y$ на $X$;

Для побудови вибіркових ліній регресії $Y$ на $X$ використовується метод найменших квадратів, для реалізації якого обирається співвідношення вигляду:

$$\begin{cases} n \cdot b_0 + b_1 \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ b_0 \cdot \sum_{i=1}^{n} x_i + b_1 \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i \cdot y_i \end{cases}$$

де $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$, $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$, $\overline{y^2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2$, $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i$, $n$ - число вимірювань.

**2** Для наочності обчислень за МНК побудуємо таблицю 46.

Таблиця 46

| $i$ | $x_i$ | $y_i$ | $x_i^2$ | $x_i y_i$ | $y_i^2$ | $\hat{y}_i$ | $e_i$ | $e_i^2$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 107 | 102 | 11449 | 10914 | 10404 | 103,63 | -1,63 | 2,66 |
| 2 | 109 | 105 | 11881 | 11445 | 11025 | 105,49 | -0,49 | 0,24 |
| 3 | 110 | 108 | 12100 | 11880 | 11664 | 106,43 | 1,57 | 2,46 |
| 4 | 113 | 110 | 12769 | 12430 | 12100 | 109,23 | 0,77 | 0,59 |
| 5 | 120 | 115 | 14400 | 13800 | 13225 | 115,77 | -0,77 | 0,59 |
| 6 | 122 | 117 | 14884 | 14274 | 13689 | 117,63 | -0,63 | 0,40 |
| 7 | 123 | 119 | 15129 | 14637 | 14161 | 118,57 | 0,43 | 0,18 |
| 8 | 128 | 125 | 16384 | 16000 | 15625 | 123,24 | 1,76 | 3,10 |
| 9 | 136 | 132 | 18496 | 17952 | 17424 | 130,71 | 1,29 | 1,66 |
| 10 | 140 | 130 | 19600 | 18200 | 16900 | 134,45 | -4,45 | 19,8 |
| 11 | 145 | 141 | 21025 | 20445 | 19881 | 139,11 | 1,89 | 3,57 |
| 12 | 150 | 144 | 22500 | 21600 | 20736 | 143,78 | 0,22 | 0,05 |
| Сума | 1503 | 1448 | 190617 | 183577 | 176834 | - | 0 | 35,3 |
| Середнє | 125,25 | 120,67 | 15884,75 | 15298,08 | 14736,17 | - | - | - |

Розв’язавши систему відносно параметрів $b_0$ і $b_1$, знайдемо:

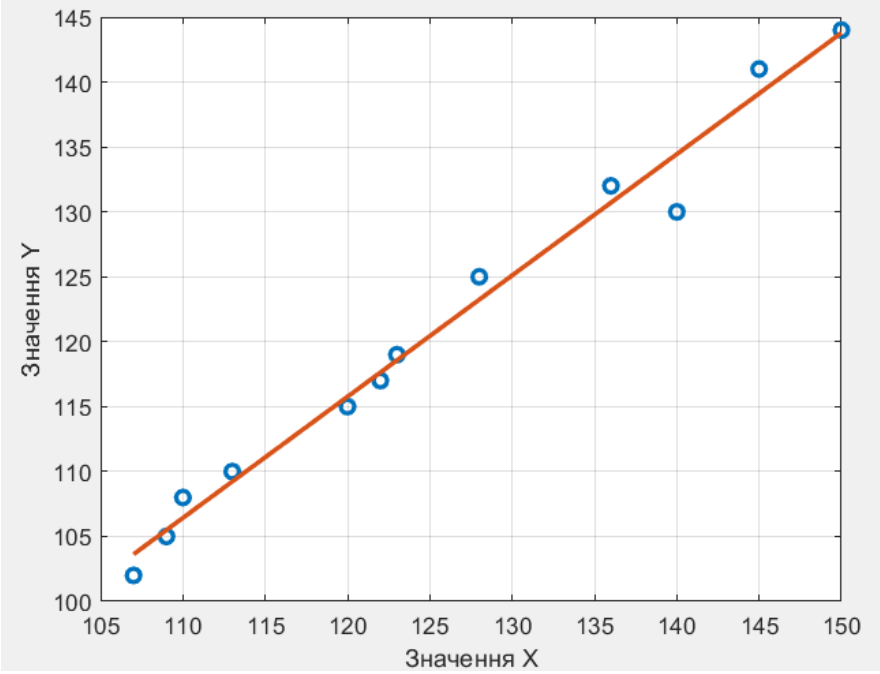
$$b_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} \cdot (\bar{x})^2} = \frac{15298,08 - 125,25 \cdot 120,67}{15884,75 - (125,25)^2} = 0,9339;$$

$$b_0 = \bar{y} - b_1 \cdot \bar{x} = 120,67 - 0,9339 \cdot 125,25 = 3,699.$$

Отже, рівняння парної лінійної регресії має наступний вигляд:

$y_i = 3,699 + 0,9339 \cdot x_i$.

4) Нанести на діаграму розсіювання графік функції регресії;



5) оцінити силу лінійної залежності між $X$ і $Y$;

$$r_{xy} = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{S_x \cdot S_y},$$

де $S_x^2 = \overline{\left(x^2\right)} - \left(\overline{x}\right)^2$, $S_y^2 = \overline{\left(y^2\right)} - \left(\overline{y}\right)^2$.

$$r_{xy} = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\sqrt{\overline{\left(x^2\right)} - \left(\overline{x}\right)^2} \cdot \sqrt{\overline{\left(y^2\right)} - \left(\overline{y}\right)^2}} =$$

$$= \frac{15298,08 - 125,25 \cdot 120,67}{\sqrt{15884,75 - \left(125,25\right)^2} \cdot \sqrt{14736,17 - \left(120,67\right)^2}} = 0,9916.$$

Знайдене значення коефіцієнта кореляції дозволяє зробити висновок про сильну лінійну залежність між змінними $X$ і $Y$.

**Оцінимо якість рівняння регресії:**

Нехай нульова гіпотеза $H_0 : b_0 = b_1 = 0$ проти альтернативної $H_1 : b_0 \neq b_1 \neq 0$.

$$F_{факт} = r_{xy}^2 \cdot \frac{n-2}{1 - r_{xy}^{\ 2}} = 0,9916^2 \cdot \frac{12-2}{1 - 0,9916^2} = 587,75.$$

За заданим рівнем значущості $\alpha = 0,05$ та по числу степенів свободи $k_1 = 1$ та $k_2 = n - 2 = 12 - 2 = 10$, за дод. 7. $F_{табл} = 4,96 \approx 5,0$. Оскільки $F_{факт} > F_{табл}$, то $H_0$ гіпотеза про випадкову природу оцінюваних характеристик відхиляється і визнається їх статистична значимість і надійність. Модель якісна.

6) оцінити статистичну значущість коефіцієнтів регресії та коефіцієнта кореляції при рівні значущості $\alpha = 0,05$.

Для цього потрібно оцінити похибки окремих параметрів та коефіцієнта кореляції згідно формул:

$$S_{r_{xy}} = \sqrt{\frac{1 - r_{xy}^{\ 2}}{n-2}} = \sqrt{\frac{1 - 0,9916^2}{12-2}} = 0,041,$$

$$S^2_{залиш} = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{35,3}{10} = 3,526;$$

$$S_{залиш} = \sqrt{S^2_{залиш}} = 1,8778$$

$$S_{b_1} = \sqrt{\frac{S^2_{залиш}}{n \cdot \sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}} = \sqrt{\frac{3,526}{12 \cdot 2366,2}} = 0,0111;$$

$$S_{b_0} = \sqrt{\frac{S^2_{залиш}}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}} = \sqrt{\frac{3,526}{2366,2}} = 0,0386.$$

Отже, спостережувані значення:

$$t^{b_1}_{спост} = \frac{b_1}{S_{b_1}} = \frac{0,9339}{0,0111} = 83,8067;$$

$$t^{b_0}_{спост} = \frac{b_0}{S_{b_0}} = \frac{3,699}{0,0386} = 95,8236;$$

$$t^{r_{xy}}_{спост} = \frac{r_{xy}}{S_{r_{xy}}} = \frac{0,9916}{0,041} = 24,1854.$$

Критичне значення для рівняння $t_{kp}\left(0,05; k = n-2 = 10\right) = 2,23$.

Нехай нульові гіпотези $H_0 : b_0 = 0; b_1 = 0; r_{xy} = 0$ та альтернативна гіпотеза $H_1 : b_0 \neq 0; b_1 \neq 0; r_{xy} \neq 0$.

Порівняємо модуль спостережуваного значення з критичним. Усі спостережувані значення більші ніж критичне, то нульові гіпотези відхиляються на користь альтернативних. Отже, це підтверджує статистичну значущість як коефіцієнтів рівняння так і кореляції.

7) розрахувати 95%-ті довірчі інтервали для теоретичних коефіцієнтів регресії;

$$b_0 - t_{kp} \cdot S_{b_0} = 3,699 - 2,23 \cdot 0,0386 = 3,6129;$$
$$b_0 + t_{kp} \cdot S_{b_0} = 3,699 + 2,23 \cdot 0,0386 = 3,7851;$$
$$b_1 - t_{kp} \cdot S_{b_1} = 0,9339 - 2,23 \cdot 0,0111 = 0,9091;$$
$$b_1 + t_{kp} \cdot S_{b_1} = 0,9339 + 2,23 \cdot 0,0111 = 0,9587.$$

Довірчий інтервал для коефіцієнта регресії береться в вигляді:

$$b_0^* \in (3,6129; 3,7851) \text{ та } b_1^* \in (0,9091; 0,9587),$$

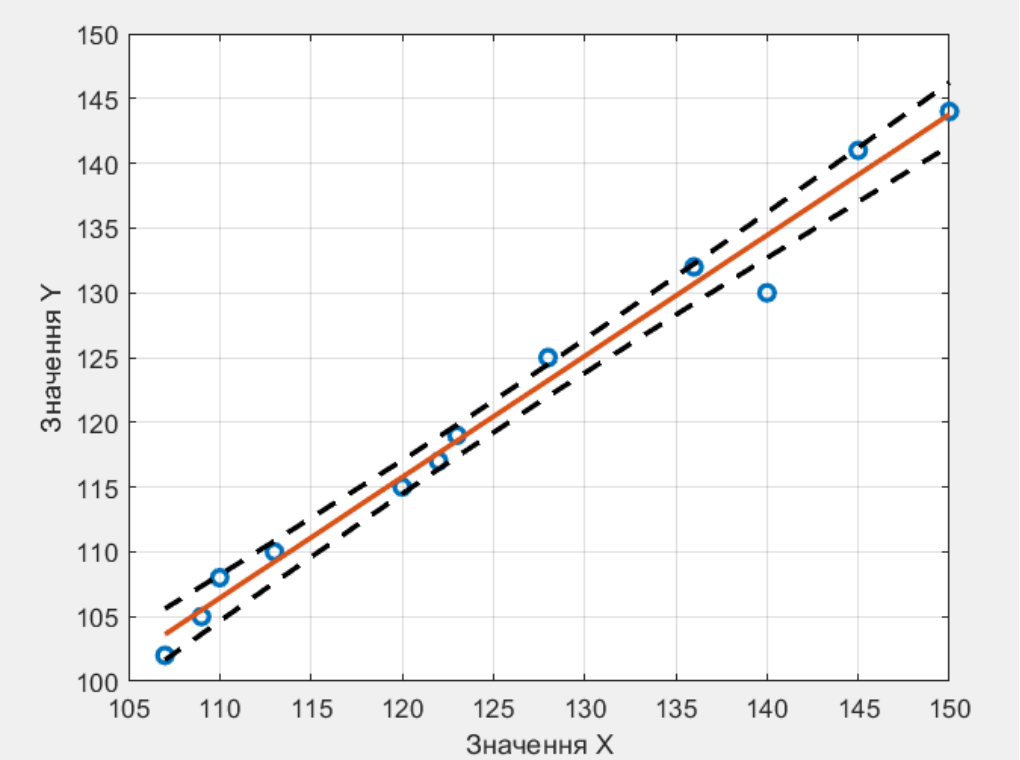
де $b_0^*$ і $b_1^*$ коефіцієнти теоретичного рівняння для всієї генеральної сукупності.

Побудуємо довірчий інтервал для лінійної парної функції регресії, а саме нижня та верхня границі довірчої області будуть знаходитися:

$$y_i^* = y_i \pm t_{табл} \cdot S_{залиш} \cdot \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} =$$

$$= y_i \pm 2,23 \cdot 1,8778 \cdot \sqrt{\frac{1}{12} + \frac{(x_i - 125,25)^2}{2366,2}};$$

| ліва межа | X | права межа |
|---|---|---|
| 101.6441 | 107.0000 | 105.6085 |
| 103.6453 | 109.0000 | 107.3429 |
| 104.6435 | 110.0000 | 108.2125 |
| 107.6256 | 113.0000 | 110.8338 |
| 114.4765 | 120.0000 | 117.0575 |
| 116.3940 | 122.0000 | 118.8756 |
| 117.3445 | 123.0000 | 119.7929 |
| 122.0064 | 128.0000 | 124.4700 |
| 129.1870 | 136.0000 | 132.2318 |
| 132.6919 | 140.0000 | 136.1981 |
| 137.0284 | 145.0000 | 141.2006 |
| 141.3344 | 150.0000 | 146.2336 |




8) спрогнозувати результативну ознаку $X = x_p$ ($x_p = 160$); визначити 95%-ті довірчі інтервали для цього прогнозу.

Прогнозоване споживання при доході у 160 складатиме

$y_{|x=160} = 3,699 + 0,9339 \cdot 160 = 153,12$.

Розрахуємо 95%-ті довірчий інтервал для умовного математичного сподівання $M\left(Y \middle| X = x_p\right)$ при $X = x_p = 160$. Скориставшись формулою

$$y_p^* = y_p \pm t_{табл} \cdot S_{залиш} \cdot \sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}} =$$

$$= 153,12 \pm 2,23 \cdot 1,8778 \cdot \sqrt{1 + \frac{1}{12} + \frac{\left(160 - 125,25\right)^2}{2366,2}} = 153,12 \pm 5,2863.$$

$y_p^* \in \left(147,8337; 158,4063\right)$.

**Множинна регресія**

Приклад 6.2. Ознака $Y$ — лінійно залежна від $x_{i1}, x_{i2}, x_{i3}$. Результати спостережень наведено в таблиці:

| $i$ | $y_i$ | $x_{i1}$ | $x_{i2}$ | $x_{i3}$ |
|---|---|---|---|---|
| 1 | 6 | 1 | 1 | 2 |
| 2 | 8 | 2 | 2 | 1 |
| 3 | 14 | 1 | 0 | 0 |
| 4 | 20 | 3 | 2 | 1 |
| 5 | 26 | 5 | 2 | 2 |

Необхідно:

1. Знайти компоненти вектора і побудувати лінійну

$$\beta^* = \begin{pmatrix} \beta_0^* \\ \beta_1^* \\ \beta_2^* \\ \beta_3^* \end{pmatrix}$$

функцію регресії $y_i = \beta_0^* + \beta_1^* x_{i1} + \beta_2^* x_{i2} + \beta_3^* x_{i3}$.

2. Обчислити коефіцієнт множинної регресії $R$.

3. Побудувати довірчий інтервал із надійністю $\gamma = 0{,}95$ для множинної лінійної функції регресії та визначення дисперсії для $\beta_0^*, \beta_1^*, \beta_2^*, \beta_3^*$ і оцінити ефективність впливу на ознаку $Y$ незалежних змінних $x_{i1}, x_{i2}, x_{i3}$.

Розв’язання

1. За умовою задачі маємо:

$$X=\begin{pmatrix}1&1&1&2\\1&2&2&1\\1&1&0&0\\1&3&2&1\\1&5&2&2\end{pmatrix},\ Y=\begin{pmatrix}6\\8\\14\\20\\26\end{pmatrix}.$$ Оскільки

$$B=\begin{pmatrix}\beta_0^*\\\beta_1^*\\\beta_2^*\\\beta_3^*\end{pmatrix}=\left(X'\cdot X\right)^{-1}\cdot X'\cdot Y=$$

$$=\frac{1}{178}\begin{pmatrix}173&-14&-39&-41\\-14&32&-38&-8\\-39&-38&123&-35\\-41&-8&-35&91\end{pmatrix}\begin{pmatrix}1&1&1&1&1\\1&2&1&3&5\\1&2&0&2&2\\2&1&0&1&2\end{pmatrix}\begin{pmatrix}6\\8\\14\\20\\26\end{pmatrix}=\begin{pmatrix}7,98\\6,34\\-3,78\\-2,58\end{pmatrix}$$

Отже, маємо $\beta_0^*=7,98;\ \beta_1^*=6,34;\ \beta_2^*=-3,78;\ \beta_3^*=-2,58$.

Рівняння регресії $\hat{y}_x=y_i^*=7,98+6,34x_{i1}-3,78x_{i2}-2,58x_{i3}+\varepsilon_i^*$.

2. Знайдемо коефіцієнт регресії $R$. Для цього скористуємося формулою

$$R=\frac{B^TX^TY-n\cdot\bar{y}^2}{Y^TY-n\cdot\bar{y}^2},$$

де $B=\begin{pmatrix}7,98\\6,34\\-3,78\\-2,58\end{pmatrix}$, $n$ - кількість спостережень; $\bar{y}$ - вибіркове середнє $y$:

$\bar{y}=\frac{\sum y_i}{n}=\frac{6+8+14+20+26}{5}=14,8$. $n\cdot\bar{y}^2=5\cdot14,8^2=1095,2$, тоді $R=0,968$.

Коефіцієнт детермінації показує, що приблизно 96,8% розкиду залежної змінної $y$ пояснюється рівнянням регресії.

3. Для побудови довірчого інтервалу для множинної лінійної функції регресії необхідно обчислити точкову незміщену квадратичну оцінку випадкового фактора $\varepsilon_i$: $S_\varepsilon^2 = S_{зал}^2 = \dfrac{\sum\left(\varepsilon_i^*\right)^2}{n-m-1}$, то в цьому разі результати обчислення зручно подати у вигляді таблиці:

| $i$ | $y_i$ | $x_{i1}$ | $x_{i2}$ | $x_{i3}$ | $y_i^* = 7{,}98 + 6{,}34x_{i1} - 3{,}78x_{i2} - 2{,}58x_{i3}$ | $y_i - y_i^*$ | $\left(\varepsilon_i^*\right)^2$ |
|---|---|---|---|---|---|---|---|
| 1 | 6 | 1 | 1 | 2 | 5,38 | 0,62 | 0,3844 |
| 2 | 8 | 2 | 2 | 1 | 10,52 | –2,52 | 6,3504 |
| 3 | 14 | 1 | 0 | 0 | 14,32 | –0,32 | 0,1024 |
| 4 | 20 | 3 | 2 | 1 | 16,86 | 3,14 | 9,8596 |
| 5 | 26 | 5 | 2 | 2 | 26,96 | –0,96 | 0,9216 |
| | | | | | | $\sum\varepsilon_i^2 = 17{,}618$ | |

Таким чином $S_{зал}^2 = \dfrac{17{,}618}{5-3-1} = 17{,}618$, тоді вибіркові дисперсії емпіричних коефіцієнтів регресії визначаємо за формулою: $S_{b_{j-1}}^2 = S_{зал}^2 \cdot z_{jj}$, $j = \overline{1,m+1}$, де $z_{jj}$ - діагональний елемент матриці $Z = \left(X^T X\right)^{-1}$, а стандартні помилки коефіцієнтів регресії визначимо як $S_{b_j} = \sqrt{S_{b_j}^2}$.

$S_{b_0}^2 = 17{,}618 \cdot \dfrac{173}{178} = 17{,}123; \quad S_{b_{залиш}} = 4{,}138;$

$S_{b_1}^2 = 17{,}618 \cdot \dfrac{32}{178} = 3{,}167; \quad S_{b_1} = 1{,}78;$

$S_{b_2}^2 = 17{,}618 \cdot \dfrac{123}{178} = 12{,}17; \quad S_{b_2} = 3{,}489;$

$S_{b_3}^2 = 17{,}618 \cdot \dfrac{91}{178} = 9{,}007; \quad S_{b3} = 3{,}001.$

Знаходимо $t$ - статистики за формулою: $t_{b_j} = \dfrac{b_j}{S_{b_j}}$, $j = \overline{0,m}$:

$$t_{b_0}=\frac{b_0}{S_{b_0}}=\frac{7,98}{4,138}=1,9285;\ t_{b_1}=\frac{b_1}{S_{b_1}}=\frac{6,34}{1,78}=3,5618;$$

$$t_{b_2}=\frac{b_2}{S_{b_2}}=-\frac{3,78}{3,489}=-1,083;\ t_{b_3}=\frac{b_3}{S_{b_3}}=-\frac{2,58}{3,001}=-0,8597.$$

За таблицею розподілу критичних точок Стьюдента визначимо критичну точку (додаток 6), де $\alpha=1-\gamma=1-0,95=0,05$.

$t_{крит}=t(\alpha,n-m-1)=t(0,05;1)=12,7$. Оскільки $|t_{b_j}|>t_{крит}$, то коефіцієнт статистично не значимий. У нашому випадку усі коефіцієнти статистично значимі, тому відповідні змінні при яких вони знаходяться не можуть бути виключені із рівняння нашої моделі.

Побудуємо 95% довірчі інтервали для коефіцієнтів множинної регресії. Для побудови застосуємо формулу $b_j-t_{крит}\cdot S_{b_j}<b_j^0<b_j+t_{крит}\cdot S_{b_j}$, $j=\overline{0,m}$.

$$b_0-t_{крит}\cdot S_{b_0}=7,98-12,7\cdot 4,138=-44,573;$$
$$b_0+t_{крит}\cdot S_{b_0}=7,98+12,7\cdot 4,138=60,533;$$
$$b_1-t_{крит}\cdot S_{b_1}=6,34-12,7\cdot 1,78=-16,266;$$
$$b_1+t_{крит}\cdot S_{b_1}=6,34+12,7\cdot 1,78=28,946;$$
$$b_2-t_{крит}\cdot S_{b_2}=-3,78-12,7\cdot 3,489=-48,0903;$$
$$b_2+t_{крит}\cdot S_{b_2}=-3,78+12,7\cdot 3,489=40,5303;$$
$$b_3-t_{крит}\cdot S_{b_3}=-2,58-12,7\cdot 3,001=-40,6927;$$
$$b_3+t_{крит}\cdot S_{b_3}=-2,58+12,7\cdot 3,001=35,5327.$$

Тоді з надійністю у 95% можна стверджувати, що коефіцієнти теоретичного рівняння належать інтервалам:

$$b_0^0\in(-44,573;\ 60,533);\quad b_1^0\in(-16,266;\ 28,946);$$
$$b_2^0\in(-48,0903;\ 40,5303);\quad b_3^0\in(-40,6927;\ 35,5327).$$

Тепер побудуємо довірчій інтервал для множинної лінійної функції регресії, для цього скористуємося формулою:

$y_i \in \left( \hat{y}_i \pm t_{крит} \cdot S_{iy} \right)$, де $S_y = \sqrt{S^2_{зал}\left(1 + x_i^T \left( X'X \right)\right)^{-1} x_i}$, $x_i = \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ x_{i3} \end{pmatrix}$,

$x_j^T = \left(1, x_{i1}, x_{i2}, x_{i3}\right)$

Якщо $x_1^T = \left(1,1,1,2\right)$, то $\hat{y}_1 \in \left(-47,3493; 58,109\right)$;

Якщо $x_2^T = \left(1,2,2,1\right)$, то $\hat{y}_2 \in \left(-32,161; 53,201\right)$;

Якщо $x_3^T = \left(1,1,0,0\right)$, то $\hat{y}_3 \in \left(-38,862; 67,502\right)$;

Якщо $x_4^T = \left(1,3,2,1\right)$, то $\hat{y}_4 \in \left(-18,444; 52,164\right)$;

Якщо $x_5^T = \left(1; 5; 2; 2\right)$, то $\hat{y}_5 \in \left(-25,006; 78,926\right)$.