Практична робота № 4

Тема: Множинний регресійний аналіз

Мета роботи: Знайти коефіцієнти регресійної залежності кількості злочинів від різних соціальних факторів. Вибрати значимі коефіцієнти та впливові фактори.

Стислі теоретичні відомості

Множинний регресійний аналіз допомагає знайти явний вигляд такої залежності та кількісно оцінити вплив різних факторів на досліджуваний процес. Але треба підкреслити, що складність розрахунків та узагальнення інформації призводить до необхідності широкого використання обчислювальної техніки, тому побудова та аналіз багатофакторних регресійних моделей базується на використанні сучасних пакетів прикладних програм.

Класична лінійна багатофакторна регресійна модель

Узагальнена багатофакторна лінійна регресійна модель може бути записана у такому вигляді:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$, де y – залежна змінна;

$x_1, x_2, \ldots, x_p$ – незалежні змінні (або фактори);

$\beta_0, \beta_1, \ldots, \beta_p$ – невідомі параметри, які потрібно оцінити;

ε – випадкова величина.

У нашому вигляді модель має p незалежних змінних, або факторів, що впливають на залежну змінну y, та $(p + 1)$ параметрів, які потрібно оцінити.

Основні припущення в багатофакторному регресійному аналізі [3].

Припущення 1

Математичне сподівання випадкової величини ε дорівнює 0:

$$M(\varepsilon_i \mid x_{1i}, x_{2i}, \ldots, x_{pi}) = 0, i = \overline{1, n}.$$

Припущення 2

Випадкові величини незалежні одна від одної, тобто відсутня серійна кореляція:

$$\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0, i \neq j.$$

Припущення 3

Модель має постійну дисперсію:

$$\mathrm{var}(\varepsilon_i) = \sigma^2.$$

Припущення 4

Коваріація між випадковою величиною ε та кожною незалежною змінною x дорівнює 0:

$$\mathrm{cov}(\varepsilon_i, x_{ji}) = 0; i = \overline{1, n}; j = \overline{1, p}.$$

Припущення 5

Модель повинна бути правильно специфікованою, тобто усі функціональні зв'язки входять у явному вигляді.

Припущення 6

Випадкова величина ε підпорядковується нормальному закону розподілу з нульовим математичним сподіванням та постійною дисперсією.

Припущення 7

Відсутність мультиколінеарності між факторами x, тобто фактори повинні бути лінійно незалежними один від одного. Не повинно бути точного лінійного зв'язку між двома або більше факторами. Припущення 7 для простої лінійної регресії відсутнє, але надзвичайно важливе для багатофакторної регресії.

Якщо всі припущення класичної лінійної регресійної моделі виконуються, то МНК-оцінки є не тільки лінійними без відхилень оцінками, але мають ще найменшу дисперсію, тобто BLUE-оцінками.

Повернемось до загальної моделі багатофакторного регресійного аналізу та знайдемо математичне сподівання обох частин. Виходячи з основних припущень, отримаємо:

$$M(y_i|x_{1i}, x_{2i},\ldots, x_{pi}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} +\ldots+ \beta_p x_{pi}.$$

Рівняння дає умовне математичне сподівання y при фіксованих значеннях x_i.

Параметри β_0, β_1,..., β_p ще називають частковими коефіцієнтами регресії. Кожний з них вимірює вплив відповідної змінної за умови, що всі інші залишаються без змін, тобто дорівнюють константам.

Етапи побудови багатофакторної регресійної моделі

Процес побудови багатофакторної регресійної моделі більш складний, ніж простої; він складається з багатьох досить кропітких етапів.

Можна виділити такі етапи побудови багатофакторної регресійної моделі:

1. Вибір та аналіз усіх можливих факторів, які впливають на процес (або показник), що вивчається.
2. Кореляційний та дисперсійний аналіз факторів.
3. Вибір методу та побудова регресійної багатофакторної моделі.
4. Оцінка невідомих параметрів моделі.
5. Перевірка моделі на адекватність.
6. Перевірка коефіцієнтів моделі на значимість.
7. Розрахунок основних характеристик та побудова інтервалів довіри.
8. Аналіз отриманих результатів, висновки.

Оцінка параметрів моделі за методом найменших квадратів (МНК)

Нехай ми маємо ряд спостережень за залежною змінною $y = \{y_1, y_2, \ldots, y_n\}$ та за незалежними змінними або факторами:

$$x_1 = \{x_{11}, x_{12}, ..., x_{1n}\}; x_2 = \{x_{21}, x_{22}, ..., x_{2n}\}; ...; x_p = \{x_{p1}, x_{p2}, ..., x_{pn}\}.$$

Виходячи з цих спостережень, побудуємо лінійну вибіркову багатофакторну модель, яку при збереженні позначень, введених для простої лінійної регресії, можна записати у вигляді:

$\widehat{y} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p + e$, де $\widehat{y}$ – залежна змінна;

$x_1, x_2, ..., x_p$ – незалежні змінні або фактори;

$b_1, b_2, ..., b_p$ – невідомі параметри;

e – випадкова величина, або помилка.

Як і у випадку простої лінійної регресії, знайдемо невідомі параметри за методом найменших квадратів, тобто мінімізуючи суму квадратів відхилень:

$$\min \sum_{i=n}^{n} e_i^2 = \sum_{i=n}^{n} (y_i - b_0 - b_1 x_1 - ... - b_p x_p)^2.$$

Для того, щоб знайти мінімум виразу, необхідно прирівняти до нуля частинні похідні. Ми отримаємо систему алгебраїчних рівнянь для знаходження коефіцієнтів.

Запис лінійної багатофакторної регресії у матричному вигляді

Лінійну багатофакторну регресійну модель, як і основні проблеми регресійного аналізу, зручно розглядати у матричному вигляді.

Запишемо модель для кожного окремого спостереження, тоді маємо:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_p,$$

де y_i – i-те значення залежної змінної $i = \overline{1, n}$;

$\beta_0, \beta_1, \ldots, \beta_p$ – невідомі параметри (істинні значення);

x_{ij} – i-те значення j-го фактора $i = \overline{1, n}$; $j = \overline{1, p}$;

ε_i – i-те значення випадкової величини.

Рівняння є скороченим записом такої системи формул:

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_p x_{1p} + \varepsilon_1; \\ y_i = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_p x_{2p} + \varepsilon_2; \\ \qquad \ldots \\ y_i = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_p x_{np} + \varepsilon_n. \end{cases}$$

$$y_{(n\times 1)} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \cdots \\ y_n \end{pmatrix}$$

– вектор-стовпець розмірності *(n X 1)* спостережень за незалежною змінною y.

$$x_{n\times(p+1)} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

– матриця розмірності (n X (p + 1)) n спостережень за p змінними $x_1, x_2, \ldots, x_p$, де перший стовпчик вміщає значення 1 для отримання перетину. Матрицю X ще називають матрицею спостережень.

$$\beta_{((p+1)\times 1)} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_p \end{pmatrix}$$ – вектор розміру (p + 1) невідомих параметрів.

$$\varepsilon_{(n\times 1)} = \begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{pmatrix}$$ – вектор розміру n випадкових величин ε_i. Виходячи з уведених позначень, отримаємо:

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}.$$

Вираз зручно переписати у матричному вигляді: $\overline{y} = X\overline{\beta} + \overline{\varepsilon}$.

Для того, щоб знайти оцінки параметрів, запишемо регресійну багато-факторну модель: $y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \dots . b_p x_{ip} + e_p$.

У матричному вигляді: $\overline{y} = X\overline{b} + \overline{\varepsilon}$.

Вектор невідомих параметрів знаходимо, мінімізуючи суму квадратів залишків:

$$e^T * e = (\overline{y} - X\overline{b})^T * (\overline{y} - X\overline{b}) = \overline{y}^T * \overline{y} - 2\overline{b}^T X^T \overline{y} + \overline{b}^T X^T X\overline{b}.$$

Прирівнявши частинні похідні до $\overline{b}$ до нуля, ми отримуємо систему рівнянь для знаходження $\overline{b}$: $X^T X\overline{b} = X^T \overline{Y}$.

Звідси $\overline{b} = (X^T X)^{-1} X^T \overline{y}$

ANOVA-дисперсійний аналіз

Елементарна ANOVA-таблиця (табл. 3) у випадку багатофакторної регресії має такий вигляд [3]:

Таблиця 4

Джерело варіації	Ступені вільності	Суми квадратів	Середні квадрати
Що пояснює регресію	$k-1$	$SSR=\sum_{i=1}^{n}(\hat{y}_i-y_{\text{ср}})^2$	$MSR = SSR/(k-1)$
Залишки	$n-k$	$SSE=\sum_{i=1}^{n}(\hat{y}_i-y_i)^2$	$MSR = SSR/(n-k)$
Загальне	$n-1$	$SST=\sum_{i=1}^{n}(y_i-y_{\text{ср}})^2$	

де k – кількість параметрів регресійної моделі, включаючи перетин.

З ANOVA-таблиці можна легко отримати вираз як для простого коефіцієнта детермінації (R^2), так і для оціненого. Ми бачимо, що у випадку багатофакторної регресії сума квадратів залишків має ($n-k$) ступенів вільності, а загальна сума квадратів – ($n-1$) ступенів вільності.

Квадрат коефіцієнта множинної кореляції називають коефіцієнтом детермінації і позначають R^2. Можна показати, що у випадку багатофакторної регресії коефіцієнт детермінації має вигляд:

$$R=\frac{\sum_{i=1}^{n}(y_i-y_{\text{ср}})(\hat{y}_i-y_{\text{ср}})}{\sqrt{\sum_{i=1}^{n}(y_i-y_{\text{ср}})^2\sum_{i=1}^{n}(\hat{y}_i-y_{\text{ср}})^2}}.$$

Оцінений, як було вказано раніше, коефіцієнт детермінації – це коефіцієнт, у якого відповідні суми квадратів скориговані на їх ступені вільності, тобто

$$R^2=1-\frac{\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}{\sum_{i=1}^{n}(y_i-y_{\text{ср}})^2}$$

$$\overline{R}^2 = 1 - \frac{SSE/(n-k)}{SST/(n-1)} .$$

Перевірка моделі на адекватність за *F*-критерієм Фішера

Для перевірки адекватності багатофакторної регресійної моделі, як і у випадку простої лінійної моделі, використовується *F*-критерій Фішера.

При цьому нуль-гіпотеза узагальнюється й має вигляд:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$

проти альтернативної гіпотези H_1: хоча б одне значення β_i, відмінне від нуля.

Якщо нуль-гіпотеза H_0 неправильна, тоді правильна гіпотеза H_1, тобто не всі параметри незначно відрізняються від нуля, що дає підставу вважати, що побудована регресійна модель відповідає дійсності, тобто адекватна. Для перевірки H_1 гіпотези розраховується *F*-статистика Фішера з $k-1$ та $(n-k)$ ступенями вільності:

$$F_{k-1,n-k} = \frac{\frac{1}{k-1} * \sum_{i=1}^{n} (\hat{y}_i - y_{\text{ср}})^2}{\frac{1}{n-k} * \sum_{i=1}^{n} (\hat{y}_i - y_i)^2},$$

де $p = k - 1$ – кількість факторів, які увійшли в модель, n – загальна кількість спостережень.

Обираємо рівень довіри $(1 - \alpha)$%). За *F*-таблицями Фішера, як і у випадку простої регресії, знаходимо критичне значення $F_{\text{кр}}$. Як і раніше, $F_{\text{кр}}$ дорівнює квантилю $(1 - \alpha)$ розподілу Фішера з $(k - 1,\ n - k)$ ступенями вільності.

Якщо $F > F_{\text{кр}}$, тоді нуль-гіпотеза відкидається, що свідчить про адекватність побудованої моделі. У протилежному випадку вона приймається, і модель вважається неадекватною.

Перевірка параметрів багатофакторної регресії на значимість

Можна показати, що у випадку багатофакторної регресії кожний параметр також розподіляється за нормальним законом розподілу: $b \sim N\,[\beta, \sigma^2 (X^T X)^{-1}]$, з математичним сподіванням, яке дорівнює істинному значенню параметра узагальненої регресії, та дисперсією, яка дорівнює дисперсії випадкової величини e^2, помноженої на відповідний діагональний елемент оберненої матриці. Але дійсне значення дисперсії випадкової величини частіше усього невідоме, тому замінюємо його на оцінку дисперсії. Така заміна приводить до того, що кожний елемент вектора буде розподілятися вже за t-розподілом Ст'юдента з $(n - k)$ ступенями вільності, що дає нам змогу обчислити t-статистику для кожного параметра:

$t = \dfrac{b_i - \beta_i}{\hat{\sigma}_{b_i}}$ зі ступенями вільності, $df = (n - k)$, де k – кількість параметрів, включаючи перетин.

Для перевірки значимості параметрів багатофакторної регресії виконуємо такі дії:

– задаємо рівень значимості $(1 - \alpha)*100\%$,
– визначаємо t-статистику для кожного параметра, яка у загальному вигляді дорівнює $t = \dfrac{b_i - \beta^*_i}{\hat{\sigma}_{b_i}}$ і підпорядкована закону розподілу Ст'юдента з $(n - k)$ ступенями вільності,

де b_i – оцінка параметра b_i, яка отримана за методом найменших квадратів;

β_i^* – гіпотетичне значення, яке має набути параметр b_i;

$\hat{\sigma}^2_{b_i}$ – оцінка дисперсій параметра b_i (з регресії);

n – кількість спостережень;

k – загальна кількість оцінених параметрів;

– висуваємо нуль-гіпотезу: H_0: $\beta_i^* = 0$ проти альтернативної H_1: $\beta_i^* \neq 0$.

У цьому разі t-статистика для параметрів має вигляд:

$$t^* = \frac{b_i}{\hat{\sigma}_{b_i}}.$$

– Значення t^*-статистики порівнюємо з критичним значенням, яке є квантиль $t_{\frac{\alpha}{2}}$ випадкової величини, підпорядкованої закону розподілу Ст'юдента з $(n - k)$ ступенями вільності, т.б. таке значення випадкової величини, яке є коренем інтегрального рівняння

$$\int_0^{t_{\frac{\alpha}{2}}} f(x)dx = \frac{(1-\alpha)}{2}$$

та відповідає межі інтервалу, для якого ймовірність влучення випадкової величини, підпорядкованої закону розподілу Ст'юдента, є $(1 - \alpha)/2$.

Якщо значення t^* потрапляє не в критичну зону, тобто $-t_{\alpha/2} < t^* < +t_{\alpha/2}$, то ми можемо зробити висновок, що з ймовірністю $(1 - \alpha)$ оцінка є статистично не значимою (тобто ми приймаємо нуль-гіпотезу, ймовірність якої велика).

Якщо значення t^* потрапляє в критичну зону, тобто $t^* < -t_{\alpha/2}$, або $t^* > t_{\alpha/2}$, то ми можемо зробити висновок, що з ймовірністю $(1 - \alpha)$ оцінка є статистично значимою (оскільки ймовірність нуль-гіпотези менше $\frac{\alpha}{2}$).

Двовимірний тест для нуль-гіпотези H_0: $\beta_i^* = 0$ показано на рис. 6 ($\alpha = 0.05$).

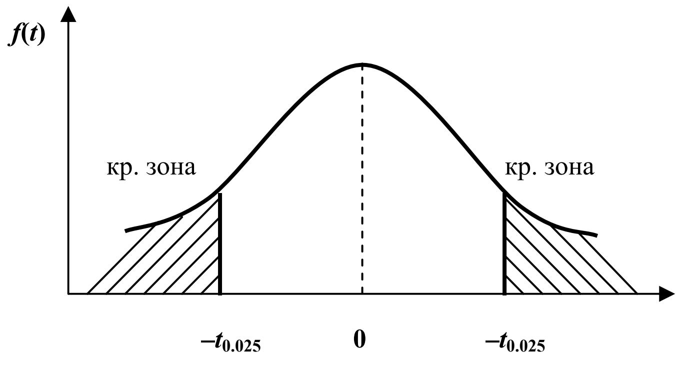


Рис. 6. Критична зона при *t*-тесті

З рисунка видно, що якщо значення t^* потрапляє в критичну (заштриховану) зону, то ми відкидаємо нуль-гіпотезу.

В багатьох програмах поряд із значенням b_i виводяться значення оцінки стандартного відхилення та відношення між b_i і цією оцінкою. Це відношення, яке є t-статистикою, називають t-значенням для відповідного параметра. Якщо воно вище від критичного значення, яке ми знаходимо за таблицею, то ми приймаємо гіпотезу H_0: $\beta_i^* = 0$ і вирішуємо, що відповідний параметр статистично значимий.

Знаходження інтервалів довіри для параметрів

Інтервальна оцінка параметрів багатофакторної регресії визначає інтервал, розташований навколо точкової оцінки b шириною $2 \cdot a$, такою, що ймовірність влучення істинного значення β у інтервал дорівнює обраному нами ступеню довіри $(1 - \alpha)$:

$$P\ \{b_i - a < \beta_i < b_i + a\} = (1 - \alpha).$$

Після алгебраїчних перетворень цей вираз приводимо до вигляду

$$P\left\{-a < \frac{b_i - \beta_i}{\hat{\sigma}_{b_i}} < a\right\} = (1 - \alpha).$$

Тут $t = \dfrac{b_i - \beta_i}{\hat{\sigma}_{b_i}}$ – випадкова величина, яка підпорядкована закону розподілу Ст'юдента з $(n - k)$ ступенями вільності.

Оскільки обираємо рівень значимості $\alpha \cdot 100\%$, відповідно рівень довіри буде дорівнювати $(1 - \alpha) \cdot 100\%$. За t-таблицею Ст'юдента знаходимо значення квантиля $\pm t_{\alpha/2}$ з $(n - k)$ ступенями вільності. Тоді ми можемо записати:

$$P(-t_{\alpha/2} < t < +t_{\alpha/2}) = (1 - \alpha).$$

Проводячи заміну $t = (b_i - \beta_i)/\hat{\sigma}_{b_i}$, отримаємо:

$$P\left\{-t_{\alpha/2} < \frac{b_i - \beta_i}{\hat{\sigma}_{b_i}} < +t_{\alpha/2}\right\} = (1 - \alpha),$$

або $P\{b_i - t_{\alpha/2}\cdot\hat{\sigma}_{b_i} < \beta_i < b_i + t_{\alpha/2}\cdot\hat{\sigma}_{b_i}\} = (1-\alpha)$.

Звідси $\beta_i = b_i \pm t_{\alpha/2}\cdot\hat{\sigma}_{b_i}$.

Прогнозування за багатофакторною регресійною моделлю

Якщо побудована регресійна модель адекватна за F-критерієм Фішера, то її можна використовувати для прогнозу залежної змінної. Припустимо, що нам відомі значення факторів у період $(n-k)$, тоді ми можемо отримати прогнозне значення $\hat{y}_{n+k}$ за моделлю
$\hat{y}_{n+k} = b_0 + b_1x_{1,n+k} + ... + b_px_{p,n+k}$, або отримаємо в матричній формі:
$\hat{y}_{n+k} = \bar{x}^T{}_{n+k} \times \bar{b}$, де $\bar{x}^T{}_{n+k} = \{1, x_{1,n+k}, \ldots, x_{p,n+k}\}$, $\bar{b}^T = \{b_0, b_1, ..., b_p\}$.

Як нам уже відомо, теорія прогнозування дає нам змогу отримати точкові та інтервальні прогнозні значення. Точкові прогнозні значення ми знаходимо за формулами:
$\hat{y}_{n+k} = b_0 + b_1x_{1,n+k} + ... + b_px_{p,n+k}$, або в матричній формі: $\hat{y}_{n+k} = \bar{x}^T{}_{n+k}\cdot b$.

Інтервальний прогноз, як і у випадку простої лінійної регресії, отримаємо для математичного сподівання залежної змінної y та для індивідуального значення y.

Для того, щоб отримати інтервальний прогноз математичного сподівання залежної змінної, розглянемо, чому дорівнює дисперсія цієї величини. Можна показати, що

$$\mathrm{var}(M(y_i(x_i))) = \mathrm{var}(\hat{y}_i(x_i)) = \sigma_\varepsilon^2 X_j{}^T(X^TX)^{-1}X_j,$$

де σ_ε^2 – дисперсія випадкової величини ε;

$X_j{}^T$ – вектор значень p факторів у момент j.

Виходячи з цього, інтервали довіри для $(1-\alpha)\cdot 100\%$ рівня довіри задля математичного сподівання y будуть дорівнювати:

$$\hat{y}_j - t_{\alpha/2}\cdot\sqrt{\hat{\sigma}^2X_j{}^T(X^TX)^{-1}X_j} \le M(y_i(x_i)) \le \hat{y}_j + t_{\alpha/2}\cdot\sqrt{\hat{\sigma}^2X_j{}^T(X^TX)^{-1}X_j}.$$

Побудова інтервалів довіри для індивідуального значення залежної змінної y.

Формула для дисперсії індивідуального значення залежної змінної y буде такою:

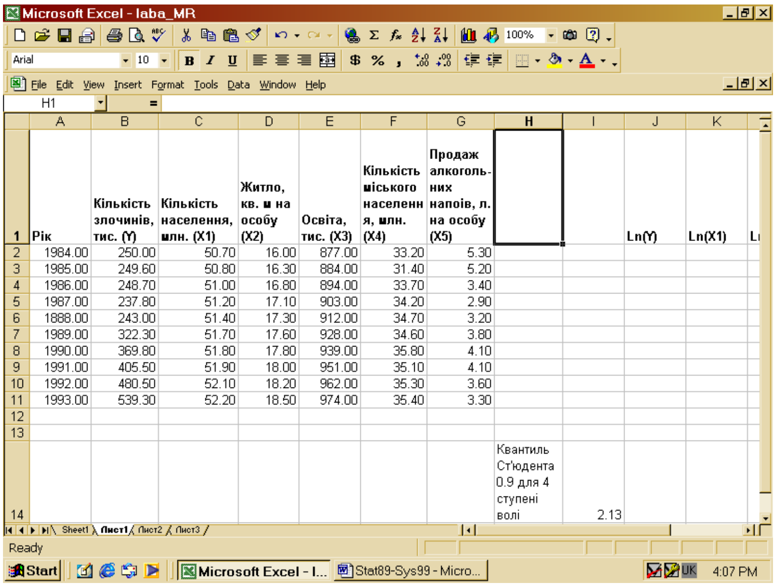
$$\operatorname{var}(y_j / x_j) = \sigma_\varepsilon^2 (1 + X_j^{\;T} (X^T X)^{-1} X_j).$$

Звідси дещо змінюється формула для обчислення інтервалу довіри, а саме:

$$\hat{y}_j - t_{\alpha/2} \cdot \sqrt{\hat{\sigma}_\varepsilon^2 * (1 + X_j^{\;T} (X^T X)^{-1} X_j)} \le y_j \le \hat{y}_j + t_{\alpha/2} \cdot \sqrt{\hat{\sigma}_\varepsilon^2 * (1 + X_j^{\;T} (X^T X)^{-1} X_j)}.$$

Приклад виконання роботи у MS Excel

Для виконання аналізу у MS Excel необхідно завантажити додатки **Analysis Tools Park** та **Analysis Tools Park VBA (Tools → Add-Ins).** Далі використовувати у головному меню пункти **Tools → DataAnalysis → Regression** (рис. 7, 8).



	A	B	C	D	E	F	G	H	I	J	K
1	Рік	Кількість злочинів, тис. (Y)	Кількість населення, млн. (X1)	Житло, кв. м на особу (X2)	Освіта, тис. (X3)	Кількість міського населення, млн. (X4)	Продаж алкогольних напоїв, л. на особу (X5)			Ln(Y)	Ln(X1)
2	1984.00	250.00	50.70	16.00	877.00	33.20	5.30				
3	1985.00	249.60	50.80	16.30	884.00	31.40	5.20				
4	1986.00	248.70	51.00	16.80	894.00	33.70	3.40				
5	1987.00	237.80	51.20	17.10	903.00	34.20	2.90				
6	1888.00	243.00	51.40	17.30	912.00	34.70	3.20				
7	1989.00	322.30	51.70	17.60	928.00	34.60	3.80				
8	1990.00	369.80	51.80	17.80	939.00	35.80	4.10				
9	1991.00	405.50	51.90	18.00	951.00	35.10	4.10				
10	1992.00	480.50	52.10	18.20	962.00	35.30	3.60				
11	1993.00	539.30	52.20	18.50	974.00	35.40	3.30				
12											
13											
14								Квантиль Ст'юдента 0.9 для 4 ступені волі	2.13		

Рис. 7

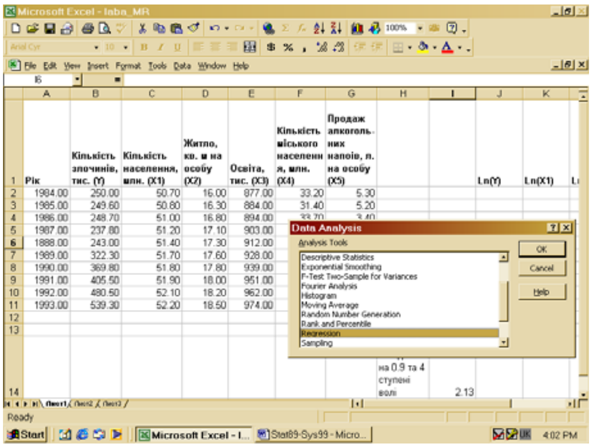


Рис. 8

Після цього у віконцях позначаєте комірки незалежних даних (*X*), комірки залежних даних (*Y*), обираєте довірчий інтервал (Confidence level – 95%), присутність сталих у рівнянні регресії, вигляд обробки помилок регресії (Constant), місце, де будуть знаходитися результати (рис. 9).

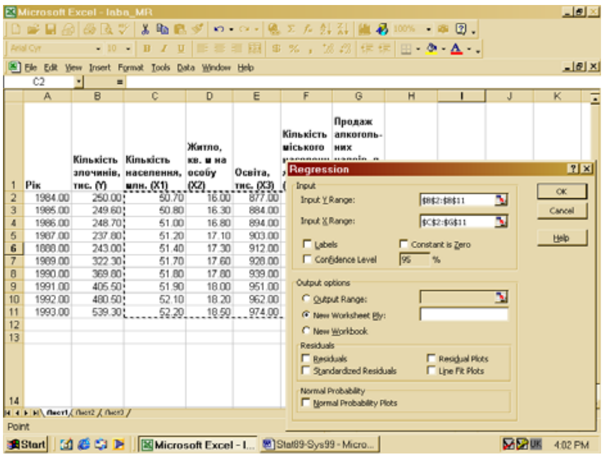
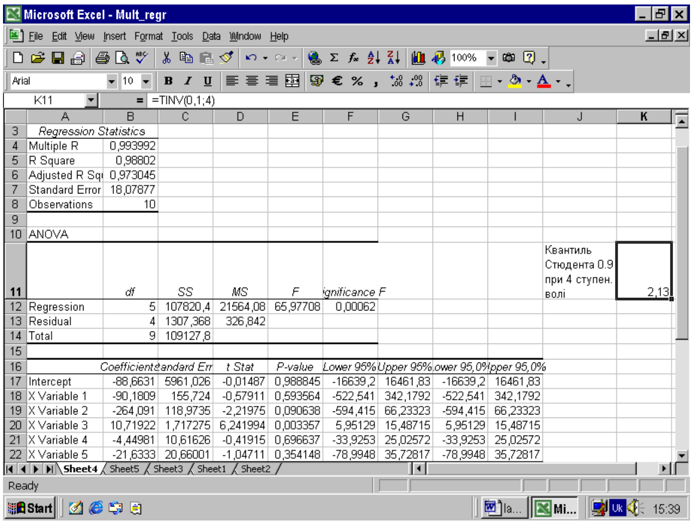


Рис. 9

Результати регресійного аналізу наведені на рис. 10, 11.



	A	B	C	D	E	F	G	H	I	J	K
3	*Regression Statistics*										
4	Multiple R	0,993992									
5	R Square	0,98802									
6	Adjusted R Squ	0,973045									
7	Standard Error	18,07877									
8	Observations	10									
9											
10	ANOVA										
11		*df*	*SS*	*MS*	*F*	*ignificance F*				Квантиль Стюдента 0.9 при 4 ступен. волі	2,13
12	Regression	5	107820,4	21564,08	65,97708	0,00062					
13	Residual	4	1307,368	326,842							
14	Total	9	109127,8								
15											
16		*Coefficients*	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95,0%*	*pper 95,0%*		
17	Intercept	-88,6631	5961,026	-0,01487	0,988845	-16639,2	16461,83	-16639,2	16461,83		
18	X Variable 1	-90,1809	155,724	-0,57911	0,593564	-522,541	342,1792	-522,541	342,1792		
19	X Variable 2	-264,091	118,9735	-2,21975	0,090638	-594,415	66,23323	-594,415	66,23323		
20	X Variable 3	10,71922	1,717275	6,241994	0,003357	5,95129	15,48715	5,95129	15,48715		
21	X Variable 4	-4,44981	10,61626	-0,41915	0,696637	-33,9253	25,02572	-33,9253	25,02572		
22	X Variable 5	-21,6333	20,66001	-1,04711	0,354148	-78,9948	35,72817	-78,9948	35,72817		

Рис. 10

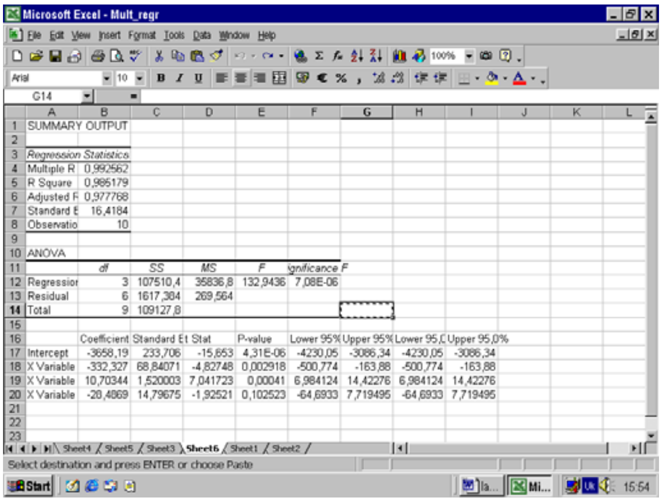
Як видно з рис. 10, критерій Фішера дорівнює 66,98, що значно перевищує квантиль 0,99 розподілу Фішера зі ступенями вільності 5 та 4, це свідчить про адекватність моделі.

Статистика Ст'юдента коефіцієнтів регресії (t Stat) не потрапляє у критичну зону |t Stat| > 2.13 для коефіцієнтів при змінних 2 та 3. Імовірність нульової гіпотези (P-value) менше 0,5 для факторів 2, 3, 5. Повторний множинний регресійний аналіз для факторів 2, 3, 5 дає такі результати (рис. 11).

Контрольні запитання

1. Яка головна ідея методу найменших квадратів для визначення коефіцієнтів регресійної залежності? Який загальний вигляд має система алгебраїчних рівнянь при визначенні коефіцієнтів регресійної залежності методом найменших квадратів?
2. Що таке багатофакторний регресійний аналіз? Головні припущення і особливі випадки.

3. Як виконується перевірка адекватності багатофакторної регресійної моделі у цілому?
4. Як перевіряються коефіцієнти регресійної моделі на значимість?
5. Як виконується знаходження інтервальних оцінок для математичного сподівання регресійної залежності та прогнозування випадкової величини?
6. Знаходження інтервальних оцінок для коефіцієнтів регресійної залежності.


Microsoft Excel - Mult_regr

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0,992562							
5	R Square	0,985179							
6	Adjusted F	0,977768							
7	Standard E	16,4184							
8	Observatio	10							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*ignificance F*			
12	Regressior	3	107510,4	35836,8	132,9436	7,08E-06			
13	Residual	6	1617,384	269,564					
14	Total	9	109127,8						
15									
16		Coefficient	Standard E	t Stat	P-value	Lower 95%	Upper 95%	Lower 95,(	Upper 95,0%
17	Intercept	-3658,19	233,706	-15,653	4,31E-06	-4230,05	-3086,34	-4230,05	-3086,34
18	X Variable	-332,327	68,84071	-4,82748	0,002918	-500,774	-163,88	-500,774	-163,88
19	X Variable	10,70344	1,520003	7,041723	0,00041	6,984124	14,42276	6,984124	14,42276
20	X Variable	-20,4869	14,79675	-1,92521	0,102523	-64,6933	7,719495	-64,6933	7,719495

Рис. 11