

## Практичне завдання №3

### Розв'язок задач лінійної регресії за допомогою надбудови “Анализ данных”

#### 1. Відомості з теорії

В аналізі економічних явищ на основі економіко-математичних методів особливе місце займають моделі, що виявляють кількісні зв'язки між *показниками*, що досліджуються, та *факторами*, які впливають на них. Предметом вивчення економетрії є вивчення кількісного аспекту економічних явищ та процесів засобами математичного та статистичного аналізу. Головним інструментом економетрії є *економетрична модель*, тобто економіко-математична модель факторного аналізу, параметри якої оцінюються засобами математичної статистики. Ця модель є засобом для аналізу та прогнозування конкретного економічного процесу на основі реальної статистичної інформації. *Регресійними* називають моделі, що засновані на рівнянні регресії або системі регресійних рівнянь. Нехай  $X=(x_1, x_2, \dots, x_p)$  – вектор незалежних змінних (факторів),  $Y=(y_1, y_2, \dots, y_m)$  – вектор залежних змінних (показників). За емпіричними даними  $(X^{(i)}, Y^{(i)})$ ,  $i = \overline{1, n}$  необхідно побудувати вектор-функцію  $f(X)$ , яка б наближено описувала зміну  $Y$  при зміні  $X$ :

$$Y \approx f(X).$$

Мається на увазі, що множина допустимих функцій, з якої підбирається  $f(X)$ , є параметричною

$$f(X)=f(X, \theta),$$

де  $\theta=(\theta_1, \theta_2, \dots, \theta_l)$  – невідомий векторний параметр. При побудові  $f(X)$  вважатимемо, що

$$Y=f(X, \theta)+\varepsilon \quad (1),$$

де перша складова – *закономірна (систематична) зміна*  $Y$  від  $X$ , а друга – випадкова величина (*похибка регресії*). Рівняння (1) і називають рівнянням регресії.

Послідовність кроків регресійного аналізу заданого економічного явища або процесу є такою.

1. Вибір  $X$  та  $Y$ , тобто змінних, які будуть використовуватись у якості незалежних та залежних (факторів та показників).
2. Вибір типу моделі, тобто класу функцій  $f(X, \theta)$ , з яких ми розшукуємо закономірну складову в рівнянні регресії.
3. Обчислення розрахункових значень  $\hat{\theta}$  невідомих параметрів  $\theta$ . Зрозуміло, що  $\hat{\theta}$  буде функцією від емпіричних даних  $(X^{(i)}, Y^{(i)})$ ,  $i = \overline{1, n}$ . Для розв'язання цієї задачі застосовують декілька методів, одним з найпоширеніших з яких є *метод найменших квадратів*.
4. Перевірка якості побудованої моделі. Під “якістю” регресійної моделі мають на увазі її адекватність та точність. *Адекватність* – це відповідність моделі процесу або об'єкту, що досліджується. Регресійна модель процесу вважається адекватною, якщо вона правильно відображає його закономірну компоненту. Це еквівалентно вимогам, щоб

остаточна компонента (похибка регресії)  $\varepsilon_i = Y^{(i)} - f(X^{(i)}, \hat{\theta})$ ,  $i = \overline{1, n}$ , задовольняла властивостям випадкової компоненти: випадковість коливань рівнів остаточної послідовності, рівність математичного сподівання випадкової компоненти нулю, відповідність розподілу випадкової компоненти нормальному закону розподілу, незалежність значень рівнів випадкової компоненти (відсутність суттєвої автокореляції). Висновок про адекватність моделі робиться, якщо перевірки всіх вказаних чотирьох критеріїв дають позитивний результат. Для адекватних моделей має сенс говорити про *точність* моделі, яка характеризується величиною відхилення розрахункових значень показників згідно рівнянню регресії від їх реальних величин. Серед показників точності назвемо середнє квадратичне відхилення, середня відносна похибка апроксимації, коефіцієнт збіжності, коефіцієнт детермінації та ін. Зрозуміло, що серед декількох адекватних моделей слід віддати перевагу тій, яка має найкращі показники точності.

##### 5. Використання побудованої моделі для аналізу та прогнозування.

Методи перевірки адекватності моделі наведемо лише для випадку  $m=1$  (одного показнику).

**Перевірка випадковості коливань рівнів остаточної послідовності.** Іншими словами, це означає перевірку гіпотези про правильність вибору типу регресії (класу функцій  $f(X, \theta)$ ). Для дослідження випадковості відхилень ми маємо послідовність значень похибки  $\varepsilon_i$ ,  $i = \overline{1, n}$  (вона є скалярною величиною).

Характер цих відхилень вивчається за допомогою декількох непараметричних критеріїв, з яких ми розглянемо лише *критерій серій*. Ряд з величин  $\varepsilon_i$  розташовують в порядку збільшення їх значень та знаходять медіану  $\varepsilon_{med}$  отриманого варіаційного ряду, тобто серединне значення для непарного  $n$  або середнє арифметичне з двох серединних значень для парного  $n$ . Повертаємось до початкової послідовності  $\varepsilon_i$  та порівнюємо значення цієї послідовності з медіаною. Якщо значення  $\varepsilon_i > \varepsilon_{med}$ , будемо ставити знак “плюс”, якщо  $\varepsilon_i < \varepsilon_{med}$  – знак “мінус”; у випадку співпадіння величин, що порівнюються, відповідним значенням  $\varepsilon_i$  нехтуємо. Таким чином, отримаємо послідовність, що складається з плюсів та мінусів, загальне число яких не перевищує  $n$ . Послідовність поруч розташованих плюсів або мінусів називають *серією*. Для того, щоб вибірка була випадковою, в ній не повинно бути надто довгих серій, а загальна кількість серій не повинна бути надто великою. Позначимо довжину найдовшої серії через  $K_{max}$ , а загальну кількість серій – через  $v$ . Вибірка вважається випадковою, якщо виконуються такі нерівності для 5%-ного рівня значущості:

$$K_{max} < [3,3(\lg n + 1)], v > 0,5[(n+1) - 1,96\sqrt{n-1}] \quad (2)$$

(квадратні дужки позначають цілу частину числа). Якщо хоча б одне з цих нерівностей порушується, то гіпотеза про випадковий характер відхилень остаточної послідовності відхиляється, і, відповідно, регресійна модель вважається неадекватною.

**Перевірка відповідності розподілу випадкової компоненти нормальному закону.** На похибку  $\varepsilon_i$  впливають багато (в більшості незалежних) факторів, як суб'єктивних (похибки вимірювань, округлення результатів і т. ін.) так і об'єктивних, що обумовлені особливостями розвитку економічних процесів. Як правило, кожний з таких факторів, розглянутий окремо, вкладає невеликий внесок в загальну величину похибки. З теорії імовірностей відомий “закон

великих чисел”, згідно з яким в таких умовах, незалежно від закону розподілу величини похибки для кожного окремого фактору, розподіл загальної похибки буде близьким до нормального. Для перевірки гіпотези про нормальний характер розподілу є розроблено декілька методів (RS-критерій, метод Вестергарда і т. ін.). Ми розглянемо критерій, побудований на основі дослідження асиметрії ( $\gamma_1$ ) та ексцесу ( $\gamma_2$ ). Вибіркові значення  $\hat{\gamma}_1$  та  $\hat{\gamma}_2$ , а також їх середньоквадратичні відхилення обчислюються за формулами

$$\hat{\gamma}_1 = \frac{\frac{1}{n} \sum_{i=1}^n \varepsilon_i^3}{\sqrt{\left( \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 \right)^3}}, \quad \sigma_{\hat{\gamma}_1} = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}},$$

$$\hat{\gamma}_2 = \frac{\frac{1}{n} \sum_{i=1}^n \varepsilon_i^4}{\left( \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 \right)^2} - 3, \quad \sigma_{\hat{\gamma}_2} = \sqrt{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}}.$$

Якщо одночасно виконуються нерівності

$$\left| \hat{\gamma}_1 \right| < 1,5\sigma_{\hat{\gamma}_1}, \quad \left| \hat{\gamma}_2 + \frac{6}{n+1} \right| < 1,5\sigma_{\hat{\gamma}_2}, \quad (3)$$

то гіпотеза про нормальний характер розподілу випадкової компоненти приймається. Якщо виконується хоча б одна з нерівностей

$$\left| \hat{\gamma}_1 \right| \geq 2\sigma_{\hat{\gamma}_1}, \quad \left| \hat{\gamma}_2 + \frac{6}{n+1} \right| \geq 2\sigma_{\hat{\gamma}_2}, \quad (4)$$

то гіпотеза про відповідність похибки нормальному розподілу відхиляється і модель визнається неадекватною. Причиною такої невідповідності може бути більш істотний вплив (порівняно з іншими) на величину похибки одного з суб'єктивних або об'єктивних факторів. В першому випадку слід більш ретельно підійти до проведення вимірювань (можливо, змінити процедуру проведення експерименту), в другому – підібрати інший тип моделі  $f(X, \theta)$ . У разі, якщо не виконуються розглянуті умови щодо нерівностей (3), (4), для перевірки нормальності розподілу є необхідно провести додаткові більш складні критерії.

**Перевірка незалежності значень рівнів випадкової компоненти** (відсутність суттєвої автокореляції) може бути проведена за декількома критеріями, найбільш популярним з яких є критерій Дарбіна-Уотсона. Розрахункове значення критерію обчислюється за формулою

$$d = \frac{\sum_{i=2}^n (\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^n \varepsilon_i^2}.$$

Якщо в результаті розрахунків ми отримаємо значення  $2 < d \leq 4$ , то в подальшому замість  $d$  необхідно використовувати величину  $d' = 4 - d$ .

Розрахункове значення критерію порівнюється з верхнім  $d_2$  та нижнім  $d_1$  критичними значеннями статистики Дарбіна-Уотсона. Якщо  $d > d_2$ , то гіпотеза про відсутність автокореляції приймається. Якщо  $d < d_1$ , то гіпотеза про відсутність автокореляції відхиляється і модель визнається неадекватною. Якщо  $d_1 \leq d \leq d_2$ , то немає достатніх підстав зробити той чи інший висновок щодо відсутності автокореляції; можливо, необхідно зробити більше спостережень.

**Перевірка рівності математичного сподівання випадкової компоненти нулю.** З курсу “Математична статистика” відомо, що величина

$$\bar{\varepsilon} = \sum_{i=1}^n \varepsilon_i$$

у випадку незалежних однаково розподілених величин  $\varepsilon_i$  є незміщеною точковою оцінкою для математичного сподівання  $\varepsilon_i$ , тобто  $M(\bar{\varepsilon}) = M(\varepsilon_i)$ ,  $i = \overline{1, n}$ .

Для  $\varepsilon_i$ , розподілених за нормальним законом, перевірка гіпотези  $M(\varepsilon_i) = 0$  здійснюється на основі критерію Стюдента, розрахункове значення якого обчислюється за формулою

$$t = \frac{\bar{\varepsilon}}{S_{\varepsilon}} \sqrt{n},$$

де

$$S_{\varepsilon} = \sqrt{\frac{\sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})^2}{n-1}}$$

- середньоквадратичне відхилення послідовності  $\varepsilon_i$ . Якщо розрахункове значення  $t$  менше табличного значення статистики Стюдента з  $n-1$  ступенями вільності, то гіпотеза про рівність математичного сподівання  $\varepsilon_i$  нулю приймається; в іншому випадку гіпотеза відхиляється і модель визнається неадекватною.

*Лінійною* називається регресійна модель

$$f(X) = a_0 + a_1 x_1 + a_2 x_2 + \dots + a_p x_p,$$

де  $a_j$ ,  $j = \overline{1, p}$  - невідомі параметри моделі. Якщо  $p=1$ , то модель називається *парною* (однофакторною, простою) *регресією*, у випадку  $p > 1$  – *множинною* (багатофакторною) *регресією*.

Розглянемо детальніше парну лінійну регресію

$$y = a_0 + a_1 x.$$

Згідно методу найменших квадратів, для розрахункових значень коефіцієнтів регресії отримаємо формули

$$\hat{a}_1 = \frac{\frac{1}{n} \sum_{i=1}^n x^i y^i - \bar{x} \bar{y}}{\frac{1}{n} \sum_{i=1}^n (x^i)^2 - \bar{x}^2}, \quad \hat{a}_0 = \bar{y} - \hat{a}_1 \bar{x},$$

де  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x^i$ ,  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y^i$ . Отже, розрахункові значення показника будуть

обчислюватись за рівнянням

$$\hat{y} = \hat{a}_0 + \hat{a}_1 x.$$

Після обчислення коефіцієнтів регресії можна приступати до оцінки якості моделі. Окрім методів такої оцінки, наведених вище, важливу роль для лінійної регресії відіграє аналіз вибіркового коефіцієнту кореляції між  $y$  та  $x$  (які розглядаються як випадкові величини)

$$\hat{r}_{xy} = \frac{\frac{1}{n} \sum_{i=1}^n x^i y^i - \bar{x} \bar{y}}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x^i)^2 - \bar{x}^2} \sqrt{\frac{1}{n} \sum_{i=1}^n (y^i)^2 - \bar{y}^2}}.$$

Квадрат вибіркового коефіцієнту кореляції  $R^2 = \hat{r}_{xy}^2$  називається *коефіцієнтом детермінації* та показує долю зміни (варіації) результативної ознаки (показника) під дією факторної ознаки та змінюється від 0 до 1. Значення  $R^2=0$  свідчить про відсутність кореляційного зв'язку між фактором та показником, тобто, іншими словами, використання рівняння регресії не дає суттєвих переваг порівняно з тривіальним вибором  $\hat{y} = \bar{y}$ . Другий крайній випадок  $R^2=1$  означає точну підгонку: всі точки спостережень лежать на регресійній прямій. Чим ближче коефіцієнт детермінації до одиниці, тим вище якість підгонки. Перевірити гіпотезу про відсутність кореляційного зв'язку можна за допомогою критерію Стюдента. Обчислюється розрахункове значення  $t$ -статистики

$$t = \hat{r}_{xy} \sqrt{\frac{n-2}{1-\hat{r}_{xy}^2}}$$

та порівнюється з табличним значенням  $t_{\beta, n-2}$ . Якщо  $|t| \geq t_{\beta, n-2}$ , то із заданою надійністю  $\beta$  гіпотезу про відсутність кореляційного зв'язку між  $x$  та  $y$  слід відкинути та прийняти альтернативну гіпотезу про наявність залежності між цими величинами.

У випадку лінійної регресії гіпотеза про відсутність кореляційного зв'язку між  $x$  та  $y$  еквівалентна гіпотезі про рівність нулю коефіцієнта  $a_1$  регресії при незалежній змінній. Для перевірки цієї гіпотези можна також скористатися критерієм Фішера (F-критерієм). Розрахункове значення цього критерію таке:

$$F_p = \frac{\hat{r}_{xy}^2 (n-2)}{1-\hat{r}_{xy}^2}$$

Якщо  $F_p > F_{(\alpha, 1, n-2)}$ , де  $F_{(\alpha, 1, n-2)}$  – табличне значення F-критерію з рівнем значущості  $\alpha$  та ступенями вільності  $(1, n-1)$ , то з надійністю  $1-\alpha$  гіпотезу  $a_1=0$  слід відкинути і прийняти альтернативну гіпотезу  $a_1 \neq 0$ .

**Прогнозування за допомогою парної лінійної регресії.** Для заданого значення незалежної змінної  $x_p$  відповідне розрахункове значення  $\hat{y}_p$  обчислюється згідно рівняння регресії

$$\hat{y}_p = \hat{a}_0 + \hat{a}_1 x_p.$$

З практичної точки зору нам бажано було б знати, в який інтервал відносно  $\hat{y}_p$  попаде реальне значення  $y$ . Радіус цього інтервалу може бути обчислений за формулою

$$\Delta y_p = t_{\alpha, n-2} S \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \quad (5)$$

де

$$S = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}}, \quad \hat{y}_i = \hat{a}_0 + \hat{a}_1 x_i, \quad i = \overline{1, n},$$

$t_{\alpha, n-2}$  – табличне значення функції Стюдента з рівнем надійної імовірності  $\alpha$  та числом ступенів вільності  $n-2$ . Отже, реальне значення  $y_p$  з надійністю  $\alpha$  попаде в інтервал  $(\hat{y}_p - \Delta y_p, \hat{y}_p + \Delta y_p)$ .

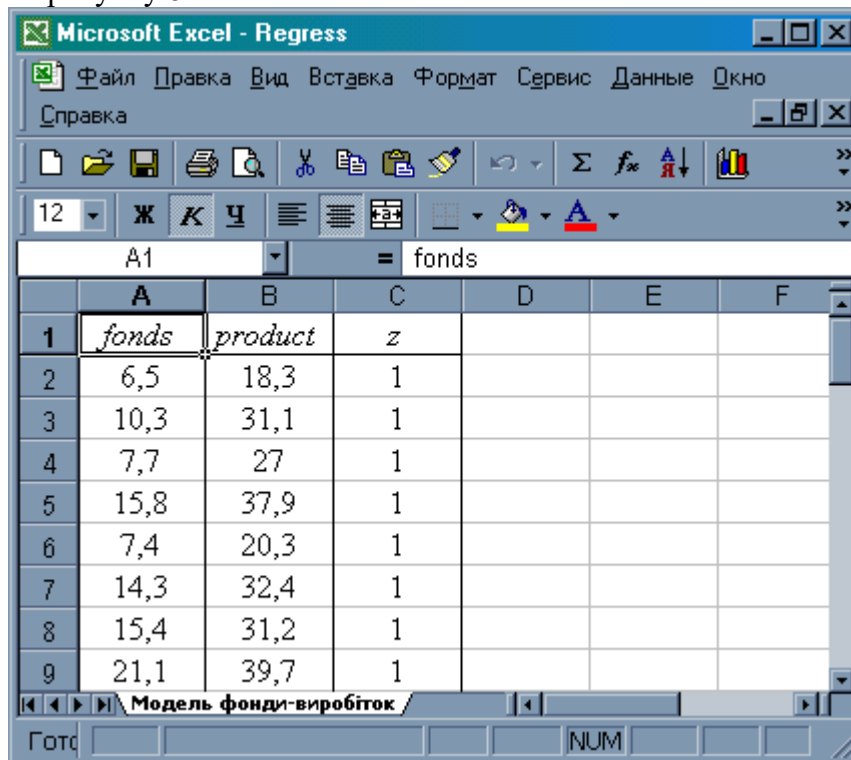
## 2. Приклад розв'язання лінійної регресійної задачі.

**Задача.** В таблиці наведені дані по 45 підприємствам легкої промисловості по статистичному зв'язку між вартістю основних фондів (*fonds*, млн руб.) і середнім виробітком на 1 працівника (*product*, тис. руб.);  $z$  - допоміжна ознака:  $z = 1$  - федеральне підпорядкування,  $z = 2$  - муніципальне.

<i>fonds</i>	<i>Product</i>	<i>z</i>	<i>fonds</i>	<i>product</i>	<i>z</i>	<i>fonds</i>	<i>product</i>	<i>Z</i>
6,5	18,3	1	9,3	17,2	2	10,4	21,4	2
10,3	31,1	1	5,7	19,0	2	10,2	23,5	2
7,7	27,0	1	12,9	24,8	2	18,0	31,1	2
15,8	37,9	1	5,1	21,5	2	13,8	43,2	2
7,4	20,3	1	3,8	14,5	2	6,0	19,5	2
14,3	32,4	1	17,1	33,7	2	11,9	42,1	2
15,4	31,2	1	8,2	19,3	2	9,4	18,1	2
21,1	39,7	1	8,1	23,9	2	13,7	31,6	2
22,1	46,6	1	11,7	28,0	2	12,0	21,3	2
12,0	33,1	1	13,0	30,9	2	11,6	26,5	2
9,5	26,9	1	15,3	27,2	2	9,1	31,6	2
8,1	24,0	1	13,5	29,9	2	6,6	12,6	2
8,4	24,2	1	10,5	34,9	2	7,6	28,4	2
15,3	33,7	1	7,3	24,4	2	9,9	22,4	2
4,3	18,5	1	13,8	37,4	2	14,7	27,7	2

Проаналізувати, чи має сенс лінійна модель зв'язку між змінними *fonds* та *product*. Чи зміняться результати аналізу, якщо обмежитись підприємствами лише федерального підпорядкування? У випадку адекватності моделі зробити прогноз виробітку на працівника при розмірі основних фондів *fonds*=25 млн. руб. для федерального підприємства. Визначити надійний інтервал прогнозу.

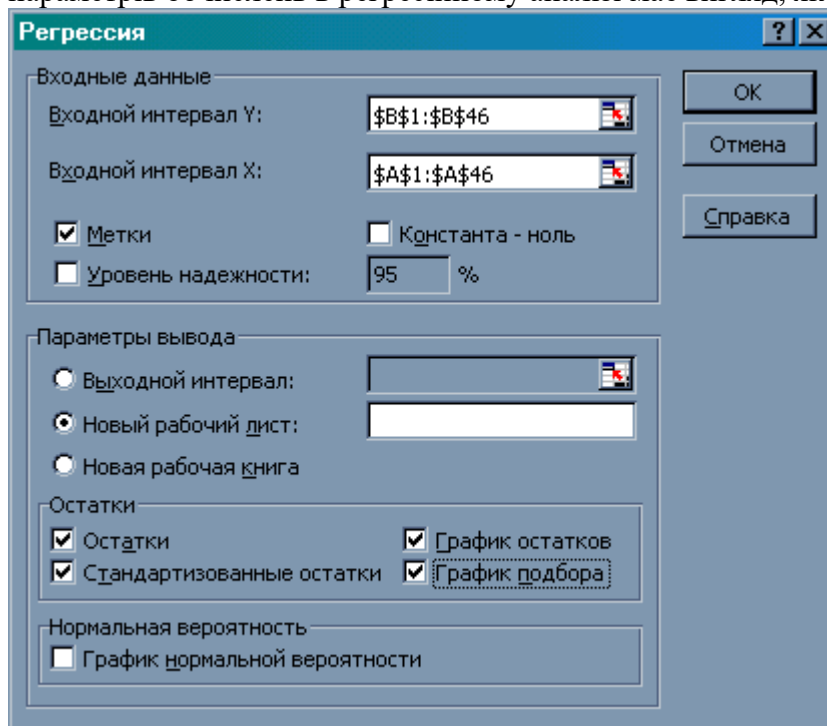
Розв'язок. В якості фактора виберемо змінну *fonds*, показника - *product*. Вхідні дані розташуємо на листі електронної таблиці в діапазоні A1:C46, як показано на рисунку 8.



	A	B	C	D	E	F
1	<i>fonds</i>	<i>product</i>	<i>z</i>			
2	6,5	18,3	1			
3	10,3	31,1	1			
4	7,7	27	1			
5	15,8	37,9	1			
6	7,4	20,3	1			
7	14,3	32,4	1			
8	15,4	31,2	1			
9	21,1	39,7	1			

Рис. 8. Вхідні дані регресійної задачі

Для регресійного аналізу вхідних даних скористуємось надбудовою “Аналіз даних” пакету Microsoft Excel. Викликаємо діалог «Анализ данных» (Сервис > Анализ данных) та вибираємо тип аналізу «Регрессия». Діалогове вікно параметрів обчислень в регресійному аналізі має вигляд, як показано на рис. 9.



**Регрессия**

Входные данные

Входной интервал Y:

Входной интервал X:

☒ Метки ☐ Константа - ноль

☐ Уровень надежности:  %

Параметры вывода

☐ Выходной интервал:

☒ Новый рабочий лист:

☐ Новая рабочая книга

Остатки

☒ Остатки ☒ График остатков

☒ Стандартизованные остатки ☒ График подбора

Нормальная вероятность

☐ График нормальной вероятности

OK Отмена Справка

Рис. 9. Діалог “Регресія”

В якості вхідного інтервалу Y вводимо діапазон для даних product (B1:B46), вхідного інтервалу X – діапазон для fonds (A1:A46). Встановлений прапорець “Метки” означає, що перший рядок в даних для X та Y використовується як мітки даних. Запрошується також, куди необхідно виводити звіт про виконанні розрахунки параметрів регресії: на новий робочий лист поточної книги (за умовчанням), в нову робочу книгу або у заданий вихідний інтервал. В останніх двох групах параметрів встановлюється необхідність включення у звіт графічних даних: графіку підбору (взаємне розташування експериментальних даних та розрахованих за регресією, графік похибок  $\varepsilon_i$  та інше).

Після завершення введення параметрів регресії запускаємо процес розрахунків. Звіт про результати розрахунків буде представлений, як у наведений нижче таблиці (сюди не включені дані про похибки регресії).

#### ВЫВОД ИТОГОВ

<u>Регрессионная статистика</u>	
Множественный R	0,772277
R-квадрат	0,596412
Нормированный R-квадрат	0,587026
Стандартная ошибка	5,008213
Наблюдения	45

#### Дисперсионный анализ

	<i>Df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Значимость F</i>
Регрессия	1	1593,83	1593,83	63,54427	5,2E-10
Остаток	43	1078,535	25,0822		
Итого	44	2672,364			

	<i>Коэффициенты</i>	<i>Стандартная ошибка</i>	<i>t-статистика</i>	<i>P-Значение</i>	<i>Верхние 95%</i>	<i>Нижние 95,0%</i>	<i>Верхние 95,0%</i>
Y-пересечение	11,50212	2,128204	5,404612	2,67E-06	7,210187	15,79405	7,210187
Fonds	1,4344	0,179942	7,971466	5,2E-10	1,071513	1,797287	1,797287

З цієї таблиці по-перше отримаємо точкову оцінку коефіцієнта кореляції  $\hat{r}_{xy} = 0,772277$  (рядок “Множественный R” в таблиці “Регрессионная статистика”), що свідчить про сильний кореляційний зв’язок між величинами X та Y.

Розраховані величини для коефіцієнтів регресії (а також оцінки дисперсії для них та довірчі інтервали) розміщені у таблиці “Коефіцієнти”. Вільний член регресії  $\hat{a}_0 = 11,50212$  розміщений у відповідному стовпці та рядку “Y-пересечение” цієї таблиці (ячейка B17 на листі “Лист1”), коефіцієнт  $\hat{a}_1 = 1,434$  – в рядку “fonds” (ячейка B18 на цьому ж листі).

Тимчасово припинимо розглядання результатів звіту про параметри регресії, зосередивши увагу на перевірці адекватності моделі. Для цього на листі вхідних даних розташуємо розрахункові значення  $\hat{y}_i$  в стовпці D (формула в ячейці D2 = Лист1!\$B\$17+Лист1!\$B\$18\*A2, протягається до ячейки D46), значення похибки регресії – у стовпці F (формула в ячейці F2 = B2-D2, протягається до ячейки F46), квадратів похибки – у стовпці G (формула в ячейці G2 = F2^2, протягається до ячейки G46), та різницю сусідніх значень похибки  $\varepsilon_i - \varepsilon_{i-1}$ , що



необхідно для перевірки критерію Дарбіна-Уотсона (формула в ячейці H3 =F3-F2, протягається до ячейки H46). Значення числа спостережень  $n$  розташуємо в ячейці E2, присвоївши їй для зручності символічне ім'я 'n' (формула =СЧЁТ(C2:C46)).

Приступаємо тепер до перевірки критерія серій (випадковості рівнів похибки  $\varepsilon_i$ ). Для цього в ячейці K2 отримаємо значення медіани виборки (формула =МЕДИАНА(\$F\$2:\$F\$46)), а в стовпці I сформуємо дані для серій – якщо відповідне значення похибки менше медіани, ставиться значення +1, якщо менше -1, якщо дорівнює медіані – значення 0 (формула в ячейці I2 =ЕСЛИ(F2>\$K\$2;ЕСЛИ(F2=\$K\$2;0;-1);1), протягається до ячейки I46). В наступному стовпці обчислюємо дані для довжини серії: якщо поточне значення в стовпці серій дорівнює попередньому, то серія продовжується, інакше починається нова серія (в ячейку J2 вносимо число 1 – нова серія починається, формула в ячейці J3 =ЕСЛИ(I2=I3;J2+1;1), протягається до ячейки J46). Таким чином, максимальна довжина серії дорівнює максимальному значенню в стовпці J (формула =МАКС(J2:J46) в ячейці K5), а загальне число серій дорівнює кількості одиниць у цьому стовпці (формула =СЧЁТЕСЛИ(J2:J46;"=1") в ячейці K8). Розрахункові значення  $K_{\max}$  та  $v$  розмістимо відповідно в ячейках L10 та L11. Перевірка критерію виконується згідно формули (2) в ячейці K13 (формула =ЕСЛИ(И(K5<L10;K8>L11);"Задовольняється";"Не задовольняється")).

Для перевірки гіпотези про нормальний розподіл похибки необхідно обчислити розрахункові значення асиметрії та ексцесу та їх дисперсій (формула =(СУММПРОИЗВ(G2:G46;F2:F46)/n)/(СРЗНАЧ(G2:G46))^1,5 для асиметрії в ячейці K16 та =КОРЕНЬ(6\*(n-2)/((n+1)\*(n+3))) в ячейці L16 для її дисперсії; формула =(СУММКВ(G2:G46)/n)/(СРЗНАЧ(G2:G46)^2)-3 для ексцесу в ячейці K19 та =КОРЕНЬ(24\*n\*(n-2)\*(n-3)/((n+1)^2\*(n+3)\*(n+5))) в ячейці L19 для його дисперсії). Перевірка критерію нормальності виконується згідно формулам (3), (4) в ячейці K22 (формула =ЕСЛИ(И(K16<1,5\*L16;K19+6/(n+1)<1,5\*L19);"Задовольняється";ЕСЛИ(И(K16>=2\*L16;K19+6/(n+1)>=2\*L19);"Не задовольняється";"Не дає відповіді"))).

Для перевірки критерію про незалежність значень рівнів похибки нам необхідно ввести формулу для розрахункового значення коефіцієнту  $d$  Дарбіна-Уотсона (формула =СУММКВ(H3:H46)/СУММ(G2:G46) в ячейці K26), а теоретичні значення коефіцієнтів  $d_1$  та  $d_2$  взяти з таблиць (див., наприклад, [1]); розмістимо їх відповідно в ячейках L26 та M26. Значення  $d'=4-d$  у випадку, якщо  $2 < d \leq 4$ , розмістимо в ячейці K28 (формула =ЕСЛИ(И(K26>2;K26<4);4-K26;K26)).

Перевірка критерію Дарбіна-Уотсона здійснюється в ячейці K31 (формула =ЕСЛИ(K28>M26;"Задовольняється";ЕСЛИ(K28<L26;"Не задовольняється";"Не дає відповіді")) в ячейці K31).

Критерій рівності математичного сподівання похибки нулю не перевіряємо, тому що застосування методу найменших квадратів для розрахунку коефіцієнтів лінійної регресії завжди дає  $\bar{\varepsilon}=0$ .

Загальний вигляд листа електронної таблиці з розрахунками наведений на рис. 10.

	C	D	E	F	G	H	I	J	K	L	M
	z	product	n			Разн.		Довжина	Медіана		
		розр		Остатки	Кв. Ост.	Ост.	Серії	серії			
1											
2	1	20,826	45	-2,525714	6,37923		1	1	-0,284591758		
3	1	26,276		4,8235677	23,26681	7,349281	-1	1			
4	1	22,547		4,4530067	19,82927	-0,37056	-1	2	Максимальна довжина серії		
5	1	34,166		3,7343698	13,94552	-0,71864	-1	3	7		
6	1	22,117		-1,816673	3,300302	-5,55104	1	1			
7	1	32,014		0,3859693	0,148972	2,202643	-1	1	Кількість серій		
8	1	33,592		-2,39187	5,721044	-2,77784	1	1	25		
9	1	41,768		-2,067948	4,276409	0,323922	1	2			
10	1	43,202		3,3976523	11,54404	5,4656	-1	1	Крозр	8	
11	1	28,715		4,3850884	19,229	0,987436	-1	2	Врозр	16	
12	1	25,129		1,7710874	3,136751	-2,614	-1	3	Критерій серій		
13	1	23,121		0,8792468	0,773075	-0,89184	-1	4	Задовольняється		
14	1	23,551		0,648927	0,421106	-0,23032	-1	5			
15	1	33,448		0,2515696	0,063287	-0,39736	-1	6	Асиметрія розр.	Похибка	
16	1	17,67		0,8299654	0,688843	0,578396	-1	7	0,591303135	0,34183011	
17	2	24,842		-7,642033	58,40066	-8,472	1	1			
18	2	19,678		-0,678194	0,459947	6,963839	1	2	Екссес розр.	Похибка	

Рис. 10. Розташування даних та обчислень на листі електронної таблиці.

Аналізуючи результати перевірки моделі на адекватність, приходимо до висновку, що критерій серій дає позитивний результат (тобто гіпотеза про випадковість значень похибки приймається), а критерії нормальності та Дарбіна-Уотсона не дають певної відповіді (для перевірки відповідних гіпотез варто застосувати інші критерії). Прийmemo, все ж таки, що модель є адекватною і повернемось до аналізу звіту про параметри лінійної регресії, що розташований на листі "Лист1". Значення коефіцієнта детермінації  $R^2=0,596412$ , яке свідчить про те, що приблизно 60% варіації показника пояснює зміною фактора *fonds*. Перевіримо гіпотезу про рівність коефіцієнта при факторній змінній нулю. Розрахункове значення F-критерію в яєйці E12 дорівнює 63,54427, а теоретичне значення з таблиці  $F_{(0,95;1;n-2)}=F_{(0,95;1;43)}\approx 4,08$ . Отже,  $F_{\text{розр}} > F_{\text{теор}}$ , і гіпотезу про рівність коефіцієнта  $a_1$  нулю відхиляємо і приймаємо альтернативну гіпотезу  $a_1 \neq 0$ . Перевірку цієї ж гіпотези можна зробити також за допомогою критерію Стюдента, вибравши його розрахункове значення  $t=7,9715$  з яєйки D18 на листі "Лист1". Відповідне теоретичне значення  $t_{\beta,n-2}$ , яке знайдемо з таблиці, дорівнює 2,021 (для  $\beta=0,95$ ). Оскільки  $t=7,9715 > t_{\beta,n-2}$ , то з надійністю 95% можна вважати справедливою альтернативну гіпотезу  $a_1 \neq 0$ . Для визначення прогнозного значення показника для значення незалежної змінної *fonds*=25, скопіюємо формулу з яєйки D46 в яєйку D48, розмістивши відповідне значення незалежної змінної в яєйці A48 (рис. 11)

The screenshot shows the 'Microsoft Excel - Regress' window. The formula bar displays  $=\text{Лист1!}\$B\$17+\text{Лист1!}\$B\$18*A48$ . The spreadsheet contains the following data:

	A	B	C	D	E	F	G	H	I	J	K
40	12	21,3	2	28,715		-7,414912	54,98091	-7,86152	1	1	
41	11,6	26,5	2	28,141		-1,641152	2,693379	5,77376	1	2	
42	9,1	31,6	2	24,555		7,0448472	49,62987	8,685999	-1	1	
43	6,6	12,6	2	20,969		-8,369154	70,04273	-15,414	1	1	
44	7,6	28,4	2	22,404		5,9964467	35,95737	14,3656	-1	1	
45	9,9	22,4	2	25,703		-3,302672	10,90765	-9,29912	1	1	
46	14,7	27,7	2	32,588		-4,887791	23,8905	-1,58512	1	2	
47	Прогноз					$t(0,95, n-2)S$	$dY_p$		Y ниж	Y верх	
48	25			47,362		2,015	5,008213	11,38387	35,97824	58,74598	
49											
50						X середнє					
51						11,075556					

Рис. 11. Розрахунки прогнозного значення показника та його надійного інтервалу.

Отримаємо  $y_p = \hat{y}_p = 47,362$ . Для встановлення надійного інтервалу цього прогнозу скористуємось формулою (5). В ячейку F48 введемо теоретичне значення  $t$ -критерію, в ячейку F51 – середнє значення  $x$  (формула  $=CP3HAЧ(A2:A46)$ ), в ячейку G48 – значення  $S$  (формула  $=КОРЕНЬ(СУММ(G2:G46)/(n-2))$ ), в ячейку H48 – значення  $\Delta y_p$  (формула  $=F48*G48*КОРЕНЬ(1+1/n+(A48-F51)^2/КВАДРОТКЛ(A2:A46))$ ). Межі надійного інтервалу прогнозу розмістимо в ячейках I48 та J48 (формули відповідно  $=D48-H48$  та  $=D48+H48$ ). Отже, з надійністю 95% можна вважати, що прогнозне значення  $y_p$  попаде в інтервал (35,97;58,75).

Проведемо аналогічні розрахунки для підприємств федерального підпорядкування ( $z=1$ ). З графіку підбору видно, що в даному випадку модель забезпечує більш щільну підгонку регресійної прямої до даних спостережень, ніж у випадку всіх підприємств. Це підтверджується також аналітичними розрахунками. Вибірковий коефіцієнт кореляції  $\hat{r}_{xy} = 0,947173$  більший, ніж в останньому випадку, і свідчить про дуже сильну кореляційну залежність між фактором та показником. Всі критерії адекватності моделі в даному випадку задовольняються, як і критерії Стюдента та Фішера про нерівність нулю коефіцієнта  $a_1$  ( $t_{розр} = 10,65 > t_{\beta, n-2} = 2,16$ ;  $F_{розр} = 113,38 > F_{теор} = F_{(0,95;1;n-2)} = F_{(0,95;1;13)} = 4,67$ ). Прогнозне значення  $\hat{y}_p(25) = 48,599$ , його надійний інтервал (41,47;55,73) значно вузькіший, ніж в попередньому випадку.

### 3. Варіанти завдань для самостійного розв'язку

Таблиця відповідності номерів варіантів порядковим номерам курсантів за журналом групи

Номер	Номер	Номер	Номер	Номер	Номер
-------	-------	-------	-------	-------	-------

курсанта в журналі групи	варіанта лабораторної роботи	курсанта в журналі групи	варіанта лабораторної роботи	курсанта в журналі групи	варіанта лабораторної роботи
1	1	10	1	19	1
2	2	11	2	20	2
3	3	12	3	21	3
4	4	13	4	22	4
5	5	14	5	23	5
6	6	15	6	24	6
7	7	16	7	25	7
8	8	17	8	26	8
9	9	18	9	27	9

### Варіант 1

Вважають, що кількість консервних банок (Y), ушкоджених при перевезеннях у товарних вагонах, є функцією швидкості вагонів (X) при поштовхах. Методом випадкового добору було обрано 13 вагонів для перевірки того, як ця гіпотеза відповідає дійсності. Виконайте перевірку і викладіть свої висновки з аргументацією. (Джерело: Дрейпер Н., Сміт Г., кн. 1, с. 82).

№	X	Y	№	X	Y	№	X	Y
1	4	27	6	3	109	11	7	168
2	3	54	7	3	28	12	3	47
3	5	86	8	4	75	13	8	52
4	8	136	9	3	53			
5	4	65	10	5	33			

### Варіант 2

Вартість експлуатації транспортних гвинтових літаків, можливо, росте зі збільшенням літного "віку" літака. Отримано такі дані: X - вік літака, Y - 6-місячна вартість експлуатації, діл.

Визначите, чи має сенс лінійна модель? Може, кращої буде інша модель?

(Джерело: Дрейпер Н., Сміт Г., кн. 1, с. 81.)

№	X	Y	№	X	Y
1	4,5	619	10	5	1194
2	4,5	1049	11	0,5	163
3	4,5	1033	12	0,5	182
4	4	495	13	6	764
5	4	723	14	6	1373
6	4	681	15	1	978
7	5	890	16	1	466
8	5	1522	17	1	549
9	5,5	987			

### Варіант 3

Хіромантия стверджує, що "лінія життя" на лівій руці людини визначає кількість років, що проживе людина. Медична наука перевіряє це за допомогою математико-статистичного аналізу.

Для перевірки проведений прямий науковий експеримент. Зібрано дані про 50 померлих:

Y - довжина "лінії життя" у сантиметрах, X - кількість прожитих років (вік).

Пропонуємо вам обробити ці дані і сформулювати науковий висновок з посиланнями на розрахункові і графічні матеріали дослідження. (Джерело: Дрейпер Н., Сміт Г. Кн. 1, с. 87-88)

X	Y	X	Y	X	Y	X	Y
19	9,75	61	7,2	68	9	75	10,2
40	9	62	7,95	69	7,8	76	6
42	9,6	62	8,85	69	10,05	77	8,85
42	9,75	65	8,25	70	10,5	80	9
47	11,25	65	8,85	71	9,15	82	9,75
49	9,45	65	9,75	71	9,45	82	10,65
50	11,25	66	8,85	71	9,45	82	13,2
54	9	66	9,15	72	9,45	83	7,95
56	7,95	66	10,02	73	8,1	86	7,95
56	12	67	9,15	74	8,85	88	9,15
57	8,1	68	7,95	74	9,6	88	9,75
57	10,2	68	8,85	75	6,45	94	9
58	8,55			75	9,75		

### Варіант 4

Експериментально був установлений вплив температури процесу дезодорації на колір кінцевого продукту. Отримано такі дані: X - температура, Y - код кольору. Побудуйте лінійну модель і оцініть її зміст. (Джерело: Дрейпер Н., Сміт Г. Кн. 1, с. 82.)

№	X	Y	№	X	Y
1	400 0,6	0,6	8	430 0,6	0,6
2	400 0,6	0,6	9	430 0,4	0,4
3	410 0,5	0,5	10	440 0,4	0,4
4	410 0,7	0,7	11	440 0,6	0,6
5	410 0,6	0,6	12	450 0,3	0,3
6	420 0,6	0,6	13	450 0,5	0,5
7	420 0,6	0,6	14	460 0,3	0,3

## Варіант 5

Японська газета "Ніхон Кейдзай Сімбун", провівши "Обстеження торгових центрів за 1984 р.", опублікувала дані про загальні обсяги продажів і площі двох груп магазинів: торгових центрів з обсягом продажів понад 20000 млн. ієн, і супермаркетів. (З метою скорочення тексту умови задачі, замінено назви конкретних японських магазинів їхніми порядковими номерами в списку даних.)

Визначите, чи існує зв'язок між площею й обсягом продажів?

У якій групі магазинів можна побудувати лінійну регресійну модель для прогнозування загального обсягу продажів?

(Адаптовано по джерелу: "Контроль якості за допомогою персональних комп'ютерів" /Т. Макіно, М. Охасі, Х. Доке, К. Макіно; перекл. з японського. М.: Машинобудування, 1991.С.34-36).

Табл.1 Торгові Центри			Табл.2 Супермаркети		
№	Продажі, Млн.. ієн	Площа, кв. м.	№	Продажі, млн. ієн	Площа, кв. м.
	Y	X		Y	X
1	219995	50555	21	18868	16368
2	214101	70514	22	16311	9000
3	198126	63755	23	16120	14920
4	197552	57198	24	15713	20722
5	158695	37490	25	15146	6964
6	157234	60607	26	15004	10798
7	144778	65532	27	14816	12742
8	106392	52205	28	14623	15400
9	98904	41079	29	14045	17577
10	95189	46046	30	13947	12650
11	91716	40997	31	13766	15000
12	88609	44650	32	13616	13461
13	86782	35331	33	13405	13714
14	86180	31787	34	13342	12051
15	83155	31768	35	13324	12355
16	78453	46840	36	12768	15698
17	77954	38498	37	12763	12688
18	62059	34204	38	12564	12300
19	61332	30938	39	12233	9096
20	57231	52200	40	12118	9589

## Варіант 6

Прогнозування динаміки цін для сценаріїв макроекономічного розвитку в 1996 р. (Адаптовано по джерелу: "Економіка України", 1997 р., № 1).

Сценарії	1. Оптимістичний		2. Реалістичний		3. Песимістичний	
Період	Індекс споживчих цін (1.01.96 = 1)	Грошова маса в обертанні (трлн. крб.)	Індекс споживчих цін (1.01.96 = 1)	Грошова маса в обертанні (трлн. крб.)	Індекс споживчих цін (1.01.96 = 1)	Грошова маса в обертанні (трлн. крб.)
Січень	<b>1</b>	833,1	<b>1</b>	833,1	<b>1</b>	833,1
Лютий	<b>1,06</b>	833,1	<b>1,105</b>	920,8	<b>1,16</b>	1027
Квітень	<b>1,19</b>	992,5	<b>1,45</b>	1212	<b>1,54</b>	1268
Червень	<b>1,29</b>	1076	<b>1,66</b>	1382	<b>2,01</b>	1654
Серпень	<b>1,39</b>	1159	<b>1,88</b>	1568	<b>2,28</b>	1816
Вересень	<b>1,49</b>	1241	<b>2,12</b>	1772	<b>2,58</b>	2140
Листопад	<b>1,53</b>	1320	<b>2,39</b>	1994	<b>2,9</b>	2354
Грудень	<b>1,68</b>	1395	<b>2,68</b>	2233	<b>3,25</b>	2618

Чи можна по цим даним дати прогноз індексу цін, і якщо так, то з яким обрієм (на скільки місяців)?

## Варіант 7

В таблиці надані значення рівня холестерину у крові хворих на гіпертонію до лікування (величина chol0) та через місяць після початку лікування (величина chol1). Побудуйте та перевірте модель для прогнозування величини chol1. Розрахуйте прогнозне значення для chol1, якщо chol0=150 та chol0=400. Яка точність цих прогнозів? (Адаптовано по джерелу «SPSS: искусство обработки информации»/А. Бююль, П. Цефель, ДиаСофт, 2002).

chol0	chol1	chol0	chol1	chol0	chol1	chol0	chol1
265	235	245	250	225	210	194	198
185	185	180	170	170	165	275	274
225	235	205	210	210	200	209	212
160	155	215	220	261	250	184	196
265	260	220	225	256	266	201	240
195	200	185	175	210	211	188	194
230	225	180	175	271	271	236	240
185	180	170	175	271	266	290	300
220	220	210	215	210	213	246	230
180	175	215	210	201	198	256	235
265	260	220	235	302	294	275	270
185	180	215	210	208	222	214	210
170	175	160	155	250	275	208	205
195	190	205	205	200	200	285	369
185	180	215	205	190	201	350	334

### Варіант 8

Середнє споживання енергії на душу населення (в тонах умовного палива – т.у.п.) та виробництво ВВП за рік для 1990р. (за даними В.Р. Хачатурова)

Країна	Споживання в т.у.п.	ВВП на душу за рік, тис. дол..
Канада	13,5	15,1
США	11	18,3
Росія	8,5	8,1
Європа	4,4	7,6
Японія	5	13,6
Індія	0,5	0,6
Китай	0,8	1,1
Азія (решта)	0,52	1,3
Африка	0,5	0,8
Австралія та Нова Зеландія	7,4	10,3
Латинська Америка	1,4	3,1
Весь світ	2,1	?

Чи існує залежність між рівнем економічного розвитку країни та рівнем споживання енергоносіїв? Висновок обґрунтуйте. Побудуйте модель та на її основі зробіть прогноз для річного показника ВВП для всього світу в цілому, маючи дані для середнього рівня споживання енергоносіїв.

### Варіант 9

Дані опитування восьми груп сімей про витрати на продукти харчування в залежності від рівня доходів родини наведені в таблиці (числа відносні в розрахунку на 100 руб. доходу та витрат)

Доходи родини (x)	1,4	3,3	5,5	7,6	9,8	12,0	14,7	18,9
Витрати на продукти харчування (y)	1,1	1,4	2,0	2,4	2,8	3,1	3,5	4,0

Побудувати лінійну одно факторну модель залежності витрат на харчування від доходів сім'ї, оцінити її адекватність та точність.