7. МЕТОДИ ПОБУДОВИ Й ДОСЛІДЖЕННЯ РЕГРЕСІЙНИХ МОДЕЛЕЙ

Завданням дослідження складних систем і процесів часто є перевірка наявності й встановлення типу зв'язку між незалежними змінними x_i (**предикторами**, **факторами**), значення яких можуть змінюватися дослідником і мають певну заздалегідь задану похибку, та залежною змінною (**відгуком**) z. Розв'язання таких завдань є предметом регресійного аналізу. Термін "Регресія" вперше був уведений Ф. Гальтоном наприкінці XIX ст. На практиці завдання регресійного аналізу зазвичай формулюють так: необхідно підібрати достатньо просту функцію, що в певному розумінні найкращим чином описує наявну сукупність емпіричних даних.

7.1. Загальна характеристика методів і задач регресійного аналізу

Класичний регресійний аналіз включає методи побудови математичних моделей досліджуваних систем, методи визначення параметрів цих моделей і перевірки їх адекватності. Він припускає, що регресія є лінійною комбінацією лінійно незалежних базисних функцій від факторів з невідомими коефіцієнтами (параметрами). Фактори й параметри є детермінованими, а відгуки – рівноточними (тобто мають однакові дисперсії) некорельованими випадковими величинами. Передбачається також, що всі змінні вимірюють у неперервних числових шкалах.

Звичайна процедура класичного регресійного аналізу є такою. Спочатку обирають гіпотетичну модель, тобто формулюють гіпотези про фактори, які суттєво впливають на досліджувану характеристику системи, і тип залежності відгуку від факторів. Потім за наявними емпіричними даними про залежність відгуку від факторів оцінюють параметри обраної моделі. Далі за статистичними критеріями перевіряють її адекватність.

При побудові регресійних моделей реальних систем і процесів вказані вище припущення виконуються не завжди. У більшості випадків їх невиконання призводить до некоректності застосування процедур класичного регресійного аналізу і потребує застосування більш складних методів аналізу емпіричних даних.

Постулат про рівноточність і некорельованість відгуків не є обов'язковим. У випадку його невиконання процедура побудови регресійної моделі певною мірою змінюється, але суттєво не ускладнюється.

Більш складною проблемою є вибір моделі та її незалежних змінних. У класичному регресійному аналізі припускають, що набір факторів задається однозначно, всі суттєві змінні наявні в моделі й немає ніяких альтернативних способів обрання факторів. На практиці це припущення не ви-

конується. Тому виникає необхідність розробки формальних та неформальних процедур перетворення й порівняння моделей. Для пошуку оптимальних формальних перетворень використовують методи факторного та дискримінантного аналізу. На сьогодні розроблено комп'ютеризовані технології послідовної побудови регресійних моделей.

Фактори в класичному регресійному аналізі вважають детермінованими, тобто вважається, що дослідник має про них всю необхідну інформацію з абсолютною точністю. На практиці це припущення часто не виконується. Відмова від детермінованості незалежних змінних зумовлює необхідність застосування моделей кореляційного аналізу. В окремих випадках можна використовувати компромісні методи **конфлюентного аналізу**, які передбачають можливість нормально розподіленого та усіченого розкиду значень факторів. Якщо ця умова виконується, побудову моделі можна звести до багаторазового розв'язування регресійної задачі.

Відмова від припущення про детермінованість параметрів моделей у регресійному аналізі призводить до суттєвих ускладнень, оскільки порушує його статистичні основи. Але на практиці це припущення виконується не завжди. У деяких випадках можна вважати параметри випадковими величинами із заданими законами розподілу. Тоді як оцінки параметрів можна брати їх умовні математичні сподівання для відгуків, що спостерігалися. Умовні розподіли та математичні сподівання розраховують за узагальненою формулою Байєса, тому відповідні методи називають **байєсівським регресійним аналізом**.

Регресійні моделі часто використовують для опису процесів, що розвиваються у часі. У певних випадках це зумовлює необхідність переходу від випадкових значень відгуків до випадкових послідовностей, випадкових процесів або випадкових полів. Однією з поширених і найпростіших моделей такого типу є **модель авторегресії**, згідно з якою відгук залежить не тільки від факторів, але також і від часу. Якщо останню залежність можна виявити, то проблема зводиться до стандартної задачі побудови регресії для модифікованого відгуку. В інших випадках необхідно використовувати більш складні прийоми.

Процедури класичного регресійного аналізу припускають, що закон розподілу відгуків є нормальним. Проте на практиці найчастішими є випадки, коли цей закон невідомий чи відомо, що він не є нормальним. Їх дослідження зумовило виникнення непараметричного регресійного аналізу, який не передбачає необхідності попереднього задання функції розподілу.

Важливою проблемою, яка виникає при оцінюванні параметрів регресійних моделей, є наявність грубих помилок серед набору аналізованих даних. Ці помилки можуть виникати внаслідок неправильних дій дослідника, збоїв у роботі апаратури, неконтрольованих короткотривалих сильних зовнішніх впливів на досліджувану систему тощо. У таких випадках

використовують два підходи, що дають змогу зменшити вплив грубих помилок на результати аналізу. У першому з них розробляють критерії та алгоритми пошуку помилкових даних. Потім ці дані відкидають. У другому підході розробляють алгоритми аналізу, які є нечутливими до наявних помилкових даних (алгоритми робастного оцінювання параметрів).

Одним з основних постулатів класичного регресійного аналізу є припущення, що найкращі оцінки параметрів можна одержати, використовуючи метод найменших квадратів. На практиці оцінки, одержані за допомогою цього методу, часто бувають недостатньо точними і містять великі похибки. Причиною цього може бути структура регресійної моделі. Якщо вона є лінійною комбінацією експонент або поліномом високого степеня, то це призводить до поганої зумовленості матриці системи нормальних рівнянь і нестійкості оцінок параметрів. Підвищення стійкості оцінок можна досягти шляхом відмови від вимоги щодо їх незміщеності. Розвиток цього напряму досліджень призвів до виникнення гребеневого, або рідж-регресійного аналізу.

Найчастіше задачу побудови регресійної моделі формулюють так. Необхідно знайти функцію заданого класу, для якої функціонал:

$$F(\alpha) = \sum_{i=1}^{n} \left(z_i(\alpha, X) - y_i \right)^2 \rightarrow \min. \qquad (7.1а)$$

У цьому виразі $z_i(\alpha, X)$ – значення функції, що апроксимує залежність, в i-ї точці, y_i – відповідне значення емпіричної залежності, α – вектор параметрів, які треба знайти, X – вектор незалежних змінних. Одержану функцію $z(\alpha, X)$ називають (**середньоквадратичною**) **регресійною моделлю**. Метод її пошуку, який базується на застосуванні критерію (7.1а), називають методом найменших квадратів.

Іноді замість функціонала (7.1а) для визначення параметрів регресійних моделей розв'язують задачі мінімізації інших функціоналів, зокрема:

$$F(\alpha) = \sum_{i=1}^{n} \left| z_i(\alpha, X) - y_i \right| \rightarrow \min; \qquad (7.1б)$$

$$F(\alpha) = \max \left| z_i(\alpha, X) - y_i \right| \rightarrow \min. \qquad (7.1в)$$

Одержувані при цьому регресійні моделі називають, відповідно, **середньоабсолютними** (**медіанними**) та **мінімаксними**. Ці моделі найчастіше використовують при побудові робастних алгоритмів регресійного аналізу, але їх практичне застосування обмежується поганою збіжністю таких алгоритмів.

Апроксимуючу функцію у випадку однієї незалежної змінної (моделі простої регресії) часто шукають у вигляді полінома $z(x) = \sum_{j=0}^{M} \alpha_j x^j$, обер-

неного полінома $z(x) = \dfrac{1}{\sum_{j=0}^{M} \alpha_j x^j}$, експоненціальних або показникових функцій $z = \alpha e^x$ чи $z = \alpha b^x$, степеневої функції $z = \alpha x^b$, лінійно-логарифмічної функції $z = \alpha_1 + \alpha_2 x + \alpha_3 1nx$, тригонометричного ряду Фур'є тощо. За наявності декількох незалежних змінних (моделі множинної регресії) найчастіше використовують функції, лінійні як за параметрами, так і за незалежними змінними $z = \alpha_0 + \sum_{i=1}^{p} \alpha_i x_i$, а також поліноміальні моделі, що є лінійними за параметрами, але нелінійними за незалежними змінними:

$$z = \alpha_0 + \sum_{i=1}^{p} \alpha_i x_i + \sum_{\substack{i,j=1 \\ i \geq j}}^{p} \alpha_{ij} x_i x_j + \sum_{\substack{i,j,k=1 \\ i \geq j \\ j \geq k}}^{p} \alpha_{ijk} x_i x_j x_k + \ldots .$$

Останні відповідають розкладу функції відгуку в ряд Тейлора. Проте можливе й використання для апроксимації інших видів залежностей.

Регресійні моделі називають **лінійними** або **нелінійними**, якщо вони є, відповідно, лінійними або нелінійними за параметрами. При цьому визначення "лінійна" часто опускають. Значення найвищого степеня предиктора в поліноміальних моделях називають **порядком моделі**. Наприклад:

$$z = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \varepsilon, \qquad (7.2)$$

де ε – похибка моделі, є лінійною моделлю третього порядку.

Вибір типу регресійної моделі є нетривіальним завданням. Для моделей, що містять одну незалежну змінну, рекомендують спочатку нанести наявні емпіричні дані на графік. Це дає можливість визначити наявність чи відсутність залежності між досліджуваними величинами, а також зробити певні припущення про тип залежності.

На рис. 7.1 як приклади наведено певні набори емпіричних точок, для яких потрібно побудувати регресійні моделі. З наведених графіків видно, що ці моделі доцільно будувати у вигляді лінійної, квадратичної та експоненціальної функцій, відповідно. Але, як правило, визначення типу моделі за графіком емпіричних даних є не настільки очевидним, тому зазвичай доводиться перевіряти декілька варіантів моделі і вибирати кращий з них за певними критеріями.

Часто як попередній етап регресійного аналізу рекомендують за допомогою методів кореляційного аналізу перевіряти наявність значущого зв'язку між досліджуваними змінними. Але при цьому слід ураховувати, що звичайні методи кореляційного аналізу дають змогу перевіряти лише гіпотезу про наявність лінійного зв'язку. Якщо зв'язок є, але він нелінійний, висновки, отримані за допомогою кореляційного аналізу, можуть бути помилковими.

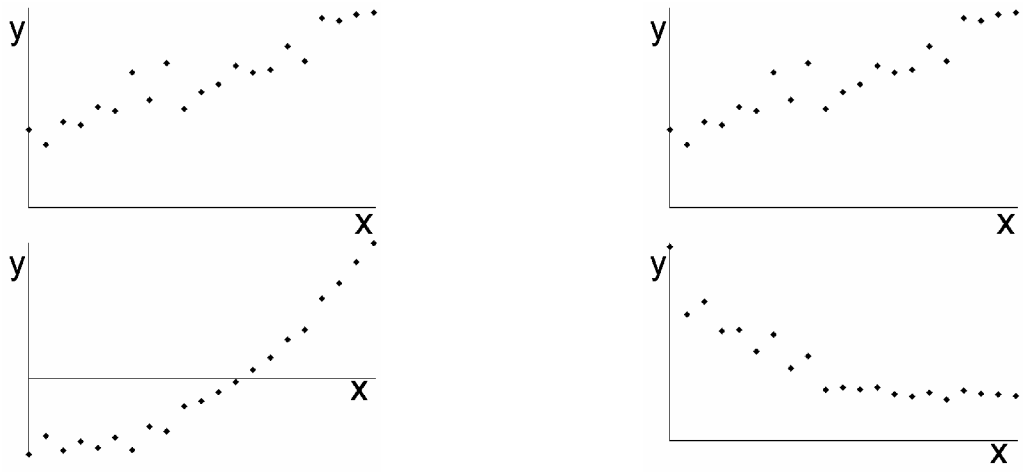


Рис. 7.1. Приклади наборів даних, для яких треба побудувати регресійні моделі

Важливою особливістю регресійних моделей є те, що їх не можна застосовувати поза межами тієї області значень вихідних параметрів, для якої вони були побудовані. При використанні регресійних моделей типу полінома, оберненого полінома, тригонометричного ряду та деяких інших слід враховувати, що, збільшуючи кількість членів ряду, можна одержати скільки завгодно близькі до нуля значення функціоналів (7.1). Проте це не завжди свідчить про якість апроксимації, оскільки ці функціонали не дають інформації про ступінь наближення моделі до емпіричної залежності у проміжках між наявними точками.

Іншою проблемою може бути наявність декількох локальних екстремумів функціоналів (7.1). У таких випадках необхідно враховувати, що більшість стандартних алгоритмів дає можливість знаходити лише локальні, а не глобальні екстремуми функціоналів, і результат мінімізації залежать від вибору початкових умов пошуку. Це часто зумовлює необхідність встановлення додаткових критеріїв вибору моделі, серед яких можуть бути як формальні критерії їх адекватності, так і неформальні критерії, що ґрунтуються на сукупності відомих даних про об’єкт дослідження.

Поліноміальні регресійні моделі, як правило, є формальними. Їх використовують для опису систем і процесів, теорію яких розроблено недостатньо. При цьому спираються на відомі властивості ряду Тейлора для аналітичних функцій. Більш цікавими для дослідників зазвичай є математичні моделі, які відображають структуру та зв’язки у системах, сутність і механізми процесів, що відбуваються у них. Якщо теоретичні основи досліджуваних систем і процесів достатньо розроблені, часто постає проблема визначення окремих параметрів моделі за наявними емпіричними даними. Для її вирішення у багатьох випадках можна використовувати формальні процедури регресійного аналізу.

На практиці часто доводиться користуватися нелінійними за параметрами та багатовимірними моделями. Під багатовимірними тут розуміють моделі, що розглядають декілька відгуків. Задачам, що розв’язуються у межах

відповідних напрямів регресійного аналізу, властиві й інші ускладнення. Так у багатовимірних моделей окремі відгуки можуть бути пов'язані один з одним. Сама регресійна модель часто задається у неявному вигляді та є неаналітичним розв'язком певної системи алгебраїчних або диференціальних рівнянь. Нестійкість оцінок параметрів для нелінійних моделей різко зростає. Як правило, такі задачі мають декілька розв'язків або не мають розв'язків взагалі.

7.2. Лінійні однофакторні моделі

Найпростішим для аналізу і найбільш дослідженим є випадок лінійної кореляційної залежності між двома змінними *X* та *Y*. Наявність лінійного зв'язку можна перевірити, розрахувавши коефіцієнт парної кореляції Пірсона (4.7).

Розглянемо детальніше задачу підбору параметрів лінійної моделі:

$$z(x) = \alpha_0 + \alpha_1 x + \varepsilon \tag{7.3}$$

за набором наявних емпіричних точок (x_i, y_i).

У **методі найменших квадратів** (МНК) виходять з припущення, що найкращими значеннями параметрів α_0 і α_1 будуть ті, для яких сума квадратів відхилень емпіричних значень y_i від розрахункових значень $z(x_i)$ набуває мінімального можливого значення. Можна довести, що МНК оцінки мають такі властивості:

– вони є лінійними функціями результатів спостережень і незміщеними оцінками параметрів моделі;

– згідно з теоремою Гауса – Маркова, МНК оцінки мають найменші дисперсії серед усіх інших оцінок, що є лінійними функціями результатів спостережень;

– МНК оцінки збігаються з оцінками, які обчислюють методом найбільшої правдоподібності.

Для знаходження таких значень параметрів необхідно розв'язати систему:

$$\begin{cases} \dfrac{\partial}{\partial \alpha_0} \sum\limits_{i=1}^{n} \left[z(x_i) - y_i \right]^2 = \dfrac{\partial}{\partial \alpha_0} \sum\limits_{i=1}^{n} \left[\alpha_1 x_i + \alpha_0 - y_i \right]^2 = 0 , \\ \dfrac{\partial}{\partial \alpha_1} \sum\limits_{i=1}^{n} \left[z(x_i) - y_i \right]^2 = \dfrac{\partial}{\partial \alpha_1} \sum\limits_{i=1}^{n} \left[\alpha_1 x_i + \alpha_0 - y_i \right]^2 = 0 . \end{cases} \tag{7.4}$$

З (7.4) можна одержати такі вирази для оцінок α^*_0 і α^*_1 коефіцієнтів лінійної залежності:

$$\alpha_1^* = \frac{n \sum\limits_{i=1}^{n} x_i y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{n \sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i \right)^2};$$

$$\alpha_0^* = \overline{Y} - \alpha_1 \overline{X} = \frac{\sum_{i=1}^{n} y_i - \alpha_1 \sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}. \qquad (7.5)$$

Перше рівняння в (7.5) є відношенням коваріації $\operatorname{cov}(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)$ до дисперсії $\sigma_X^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2$, тобто:

$$\alpha_1^* = \frac{\operatorname{cov}(x,y)}{\sigma_X^2} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}. \qquad (7.6)$$

У багатьох випадках, завдяки особливостям округлення чисел у ЕОМ, останній вираз дає змогу отримати точніші оцінки параметрів, ніж (7.5).

У випадку однофакторної лінійної моделі існує зв'язок між коефіцієнтом a_1 моделі, коефіцієнтом кореляції предиктора і відгуку, а також їх дисперсіями:

$$r_{xy} = a_1 \frac{\sigma_x}{\sigma_y}. \qquad (7.7)$$

Як приклад розглянемо таку задачу. У табл. 7.1 подано емпіричні дані, для яких треба побудувати регресійну модель, а також дані, необхідні для розрахунку її параметрів. На рис. 7.2 наведено емпіричні точки та графік залежності, побудований за одержаною методом найменших квадратів моделлю. Як видно з рис. 7.2, одержана модель задовільно описує наявні емпіричні дані.

Таблиця 7.1

Приклад розрахунку регресійної моделі

№ випробування	x_i	y_i	x_i^2	$x_i \cdot y_i$	z	Залишок
1	0	0,310	0	0	0,516	0,206
2	0,3	1,037	0,09	0,311	1,506	0,470
3	0,6	2,513	0,36	1,508	2,497	−0,017
4	0,9	3,843	0,81	3,459	3,487	−0,356
5	1,2	4,840	1,44	5,807	4,477	−0,363
6	1,5	6,020	2,25	9,030	5,467	−0,553
7	1,8	5,865	3,24	10,557	6,457	0,592
8	2,1	7,470	4,41	15,686	7,447	−0,022
9	2,4	8,889	5,76	21,332	8,438	−0,451
10	2,7	9,25399	7,29	24,98577	9,427681	0,174

Продовження табл. 7.1

11	3	10,39294	9	31,17882	10,41785	0,025
12	3,3	11,11287	10,89	36,67247	11,40801	0,295
Σ	19,8	71,54526	45,54	160,5277		0
α_1^*	3,300548	α_0^*	0,516201			

Для розглянутої регресійної моделі сума залишків $\sum_{i=1}^{n}(y_i - z(x_i))$ дорівнює нулю, якщо модель містить вільний член α_0. Виключення вільного члена з моделі зазвичай є невиправданим. Використання моделі з $\alpha_0 = 0$ доцільно лише у випадках, коли з теорії відомо, що для нульових значень предикторів відгук має дорівнювати нулю. Якщо це невідомо, але бажано одержати модель, що не містить вільного члена, більш доцільним є застосування центрування даних.

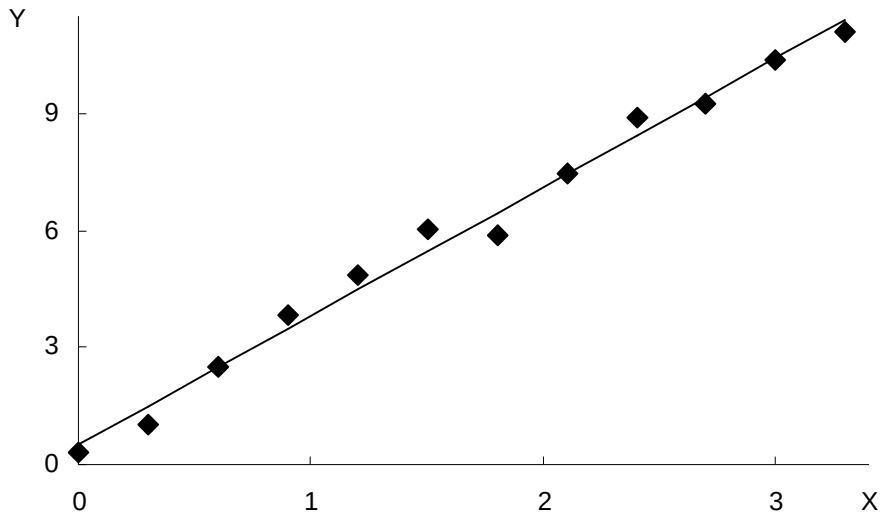


Рис. 7.2. Графіки досліджуваної залежності й лінійної моделі

Підставляючи до моделі $z(x) = \alpha_0 + \alpha_1 x + \varepsilon$ оцінку коефіцієнта α^*_0 з (7.5), можна одержати:

$$Y^* = \alpha_0^* + \alpha_1^* x + \varepsilon = \overline{Y} + \alpha_1^*\left(x - \overline{X}\right) + \varepsilon. \qquad (7.8)$$

Звідси отримуємо **центровану модель**:

$$Y - \overline{Y} = \alpha_1^*\left(x - \overline{X}\right) + \varepsilon, \qquad (7.9)$$

яка не містить вільного члена.

Незважаючи на те, що, як правило, реальні залежності відгуків від факторів є нелінійними, розглянутий випадок широко використовують у практиці побудови регресійних моделей. Це пов'язано з трьома основними причинами. По-перше, він є найбільш простим і дослідженим. Зокрема,

для нього достатньо повно розроблені процедури визначення статистичних характеристик одержуваних оцінок параметрів (дисперсії, довірчих інтервалів тощо) та перевірки адекватності моделей. По-друге, у багатьох випадках складні залежності можна подати як набір лінійних (на малих відрізках змінювання факторів) залежностей. По-третє, нелінійні залежності у деяких випадках можна перетворити до лінійного вигляду шляхом заміни змінних. Деякі приклади такого перетворення наведено у табл. 7.2.

Таблиця 7.2

Приклади лінеаризації нелінійних залежностей

Вихідна залежність	Лінеаризована залежність	Нові змінні
$z = \alpha_0 exp\,(-\alpha_1 x)$	$\ln z = \ln \alpha_0 - \alpha_1 x$	x, $\ln z$
$z = \alpha_0[1 - exp(-\alpha_1 x)]$	$\ln \frac{\alpha_0}{\alpha_0 - z} = \alpha_1 x$	x, $\ln \frac{\alpha_0}{\alpha_0 - z}$
$z = \alpha_0 exp(-\alpha_1/x)$	$\ln z = \ln \alpha_0 - \alpha_1/x$	$1/x$, $\ln z$
$z = z=\alpha_0 x^{\alpha_1}$	$\ln z = \ln \alpha_0 + \alpha_1 \ln x$	$\ln x$, $\ln z$
$z = \alpha_0 x + \alpha_1 x^2$	$z/x = \alpha_0 + \alpha_1 x$	x, z/x
$z = \alpha_0 \sin(\alpha_1 x)$	$\arcsin (z/\alpha_0) = \alpha_1 x$	x, $\arcsin (z/\alpha_0)$

Перетворення нелінійних залежностей до лінійних є строго обґрунтованим, якщо вихідні дані є точними. На практиці вони завжди вимірюються з деякою похибкою. Розглянемо модель:

$$z = \alpha_0 x^{\alpha_1} + \varepsilon, \tag{7.10}$$

де ε – похибка вимірювань.

Її лінеаризована форма матиме вигляд:

$$\ln z = \ln \alpha_0 + \alpha_1 \ln x + \varepsilon', \tag{7.11}$$

де ε' є невідомою випадковою величиною. Використання як лінеаризованої форми виразу:

$$\ln z = \ln \alpha_0 + \alpha_1 \ln x \tag{7.12}$$

є коректним лише у тому випадку, коли величина ε' є малою порівняно з іншими доданками правої частини (7.12).

Розглянемо питання про точність оцінок. Для цього запишемо таку тотожність:

$$\left(y_i - \overline{Y}\right) = \left(y_i^* - \overline{Y}\right) + \left(y_i - y_i^*\right). \tag{7.13}$$

Тут y^*_i є оцінкою значення величини у при $x = x_i$. Якщо піднести обидві частини цієї тотожності до квадрата та взяти суму від $i = 1$ до n, то одержимо:

$$\sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(y_i^* - \overline{Y}\right)^2 + \sum_{i=1}^{n}\left(y_i - y_i^*\right)^2. \tag{7.14}$$

У цієї рівності немає члена $2\sum_{i=1}^{n}\left(y_i^* - \overline{Y}\right)\left(y_i - y_i^*\right)$, оскільки:

$$y_i - \overline{Y} = \alpha_1\left(x_i - \overline{X}\right);$$

$$y_i - y_i^* = y_i - \overline{Y} - \alpha_1\left(x_i - \overline{X}\right);$$

$$\sum_{i=1}^{n}\left(y_i - \overline{Y}\right)\left(y_i - y_i^*\right) = \alpha_1\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right) - \alpha_1^2\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2 = 0,$$

враховуючи (7.6).

Розглянемо склад виразу (7.14). Його ліва частина є сумою квадратів відхилень значень, що спостерігалися, стосовно загального середнього. Перший доданок правої частини є сумою квадратів відхилень оцінок цих значень, зроблених на основі обраної моделі регресії, від загального середнього. Її часто називають сумою квадратів стосовно регресії. Другий доданок правої частини є сумою квадратів відхилень значень, що спостерігалися, від їх оцінок, одержаних з використанням обраної моделі. Цей доданок називають сумою квадратів, що зумовлена регресією. Для того, щоб модель була придатною для прогнозування значень досліджуваної величини, необхідно, щоб він був малим порівняно із сумою квадратів стосовно регресії. У граничному випадку він має дорівнювати нулю.

Будемо вважати **дисперсію залишків** $\sigma_{\varepsilon_i}^2$ і, відповідно, **дисперсію відгуків** $\sigma_{Y_i}^2$ сталими. **Дисперсія** емпіричних точок **стосовно середнього** $\sigma_{Y_i}^2$ буде дорівнювати їх **дисперсії** σ_{YX}^2 **стосовно лінії регресії** у випадку, коли постульована модель є істинною. У протилежному випадку $\sigma_{Y_i}^2 > \sigma_{YX}^2$. Оцінкою величини σ_{YX}^2 є відношення суми квадратів відхилень спостережень від середнього до кількості степенів вільності. Останню беруть рівною різниці між кількістю випробувань і кількістю констант, які визначаються незалежно одна від одної за їх результатами. У випадку, що розглядається, воно дорівнює $n - 2$, оскільки з емпіричних даних визначають два параметри прямої регресії. Тобто:

$$\sigma_{XY}^2 = \frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}{n-2}. \qquad (7.15)$$

Для прикладу, що наведений у таблиці 7.1, можна одержати $\sigma_{YX}^2 = 14,02$, $\sigma_{Y_i}^2 = 14,17$, $\sigma_{\varepsilon_i}^2 = 0,15$. Таким чином рівність $\sigma_{Y_i}^2 = \sigma_{YX}^2 + \sigma_{\varepsilon_i}^2$, яка випливає з (7.14), є виконаною.

Нехай p_i є кількістю повторних вимірювань величини $\overline{Y_i}$ при заданому значенні x_i. Тоді квадратична форма, яку мінімізують в методі найменших квадратів:

$$Q = \sum_{i=1}^{n} \left(\overline{Y_i} - z(x_i) \right)^2 p_i = \sum_{i=1}^{n} p_i \left[\overline{Y_i} - \alpha_0 - \alpha_1 \left(x_i - \overline{x} \right) \right]^2 . \qquad (7.16)$$

Розглядаючи її як функцію параметрів α_0, α_1, як і у попередньому випадку, одержуємо оцінки параметрів:

$$\begin{cases} \alpha_0^* = \sum_{i=1}^{n} p_i \overline{Y_i} / \sum_{i=1}^{n} p_i = \overline{Y} ; \\ \alpha_1^* = \sum_{i=1}^{n} p_i \overline{Y_i} \left(x_i - \overline{x} \right) / \sum_{i=1}^{n} p_i \left(x_i - \overline{x} \right)^2 . \end{cases} \qquad (7.17)$$

У припущенні, що умовний розподіл величини $\overline{Y_i}$ при заданому x_i є нормальним, оцінкою дисперсії буде величина:

$$\sigma_{Y_i}^{2\,*} = \frac{1}{n} \sum_{i=1}^{n} p_i \left(\overline{Y_i} - Y_i^* \right)^2 , \qquad (7.18)$$

де $Y_i^* = \alpha_0^* + \alpha_1^* (x_i - \overline{x_i})$.

Слід зазначити, що висновки, одержувані на основі мінімізації дисперсії похибки, є правильними тільки тоді, коли постульована модель коректна. В інших випадках вони можуть виявитися помилковими. Перед прийняттям рішення стосовно моделі треба перевірити гіпотезу, що лінійна модель $z = \alpha_0 + \alpha_1 \left(x - \overline{x} \right)$ задовільно описує емпіричні дані із заданою точністю при заданому рівні значущості η. Для цього визначають міру похибки емпіричних даних:

$$S_a^2 = \frac{1}{n-2} \sum_{i=1}^{n} p_i \left(\overline{Y_i} - Y_i^* \right)^2 . \qquad (7.19)$$

Ця величина є зміщеною оцінкою дисперсії $\sigma_{Y_i}^2$. Іноді її називають **дисперсією неадекватності**.

Незміщеною оцінкою цієї дисперсії є величина:

$$S_e^2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{p_i} \left(Y_{ij} - \overline{Y_i} \right)^2}{\sum_{i=1}^{n} p_i - n}, \qquad (7.20)$$

де $Y_{ij} - j$-те одиничне вимірювання при $x = x_i$.

Критерієм адекватності моделі при заданій надійності $1-\eta$ є виконання нерівності:

$$S_a^2 / S_e^2 \le F_{1-\eta}, \tag{7.21}$$

де $F_{1-\eta}$ – відповідне значення функції розподілу Фішера, для кількостей степенів вільності $n_1 = n_2 = n - 1$.

Іноді вважають, що малі значення відношення (7.21) свідчать про адекватність обраної моделі. Але такий висновок може виявитися помилковим. Більш докладно це питання буде розглянуто нижче.

Довірчі інтервали для параметрів α_0, α_1 можна знайти за допомогою коефіцієнтів *t*-розподілу Стьюдента з кількістю степенів вільності $\sum_{i=1}^{n} p_i - 2$:

$$\begin{cases} \alpha^* - t_{1-\eta/2} S_{\alpha^*} \le \alpha \le \alpha^* - t_{1-\eta/2} S_{\alpha^*}; \\ S_{\alpha^*} = \sqrt{\dfrac{1}{n-2} \sum_{i=1}^{n} p_i \left(\overline{\alpha} - \alpha^*\right)^2}. \end{cases} \tag{7.22}$$

Важливим практичним завданням є перевірка гіпотези про збіг двох рівнянь регресії:

$$z_1(x) = \alpha_{01} + \alpha_{11} x;$$

та

$$z_2(x) = \alpha_{02} + \alpha_{12} x.$$

Воно передбачає перевірку трьох простих гіпотез. Спочатку перевіряють гіпотезу про рівність дисперсій неадекватності моделей:

$$H_0^{(1)} : \sigma_{a1}^2 = \sigma_{a2}^2. \tag{7.23}$$

Для цього використовують дисперсійний критерій Фішера.

Якщо різниця дисперсій неадекватності є незначущою, то переходять до перевірки гіпотези про рівність кутових коефіцієнтів моделей:

$$H_0^{(2)} : \alpha_{11} = \alpha_{12}. \tag{7.24}$$

$$t_a = \frac{a_{11} - a_{12}}{\sigma \sqrt{\dfrac{1}{\sum_{i=1}^{k_1} n_{1i} \left(x_i^{(1)} - \overline{x}^{(1)}\right)^2} + \dfrac{1}{\sum_{i=1}^{k_2} n_{2i} \left(x_i^{(2)} - \overline{x}^{(2)}\right)^2}}}, \tag{7.25}$$

При цьому виходять з того, що величина підпорядковується *t*-розподілу Стьюдента з $N_1 + N_2 - 4$ степенями вільності. У (7.25) n_{1i}, n_{2i} – кількість вимірювань для першої та другою моделі в *i*-ї точці; k_1, k_2 – кількість точок для кожної з моделей;

$$\sigma = \frac{(N_1 - 2)\sigma_{a1}^2 + (N_2 - 2)\sigma_{a2}^2}{N_1 + N_2 - 4}; \tag{7.26}$$

$$N_1 = \sum_{i=1}^{k_1} n_{1i}\ ;\ N_2 = \sum_{i=1}^{k_2} n_{2i}\ .$$

Справедливість гіпотези $H_0^{(2)}$ означає, що порівнювані лінії регресії паралельні одна одній. У цьому випадку можна отримати уточнену оцінку коефіцієнта нахилу прямою регресії:

$$\overline{a_1} = \frac{\sum_{i=1}^{k_1} n_{1i}\left(x_i^{(1)} - \overline{x}^{(1)}\right)\left(y_i^{(1)} - \overline{y}^{(1)}\right) + \sum_{i=1}^{k_2} n_{2i}\left(x_i^{(2)} - \overline{x}^{(2)}\right)\left(y_i^{(2)} - \overline{y}^{(2)}\right)}{\sum_{i=1}^{k_1} n_{1i}\left(x_i^{(1)} - \overline{x}^{(1)}\right)^2 + \sum_{i=1}^{k_2} n_{2i}\left(x_i^{(2)} - \overline{x}^{(2)}\right)^2}. \tag{7.27}$$

Після цього необхідно перевірити останню гіпотезу про рівність вільних членів моделей:

$$H_0^{(3)} : a_{01} = a_{02}. \tag{7.28}$$

З цією метою використовують те, що величина

$$u = \frac{\mathbf{a}_1 - \overline{a_1}}{\sigma\left\{\mathbf{a}_1 - \overline{a_1}\right\}}, \tag{7.29}$$

де

$$\mathbf{a}_1 = \frac{\overline{y}^{(1)} - \overline{y}^{(2)}}{\overline{x}^{(1)} - \overline{x}^{(2)}},$$

$$\sigma\left\{\mathbf{a}_1 - \overline{a_1}\right\} =$$

$$= \sigma^2\left[\frac{1}{\left(\overline{x}^{(2)} - \overline{x}^{(1)}\right)^2}\left(\frac{1}{N_1} + \frac{1}{N_2}\right) + \frac{1}{\sum_{i=1}^{k_1} n_{1i}\left(x_i^{(1)} - \overline{x}^{(1)}\right)^2 + \sum_{i=1}^{k_2} n_{2i}\left(x_i^{(2)} - \overline{x}^{(2)}\right)^2}\right]$$

підпорядковується *t*-розподілу Стьюдента з $N_1 + N_2 - 4$ степенями вільності.

7.3. Поліноміальні моделі

У багатьох випадках емпіричні залежності можна описати поліноміальними моделями вигляду:

$$z = \sum_{i=1}^{q} \alpha_i x^i\ . \tag{7.30}$$

Оцінки параметрів таких моделей отримують шляхом розв'язування нормальних рівнянь виду:

$$\begin{pmatrix} n & \sum x_i & \sum x_i^2 & \dots & \sum x_i^q \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \dots & \sum x_i^{q+1} \\ \dots & \dots & \dots & \dots & \dots \\ \sum x_i^q & \sum x_i^{q+1} & \sum x_i^{q+2} & \dots & x_i^{2q} \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \dots \\ \alpha_q \end{pmatrix} = \begin{pmatrix} \sum Y_i \\ \sum Y_i x_i \\ \dots \\ \sum Y_i x_i^q \end{pmatrix}. \quad (7.31)$$

Зазвичай стовпці, що утворюють матрицю *X*, не є ортогональними. У зв'язку з цим у разі необхідності збільшення степеня полінома необхідно перераховувати оцінки всіх його коефіцієнтів. Тому для поліномів високих степенів більш раціональним методом побудови регресійної моделі є заміна вихідного рівняння (7.30) іншим:

$$z = \sum_{i=1}^{q} \alpha_i' \zeta_i , \quad (7.32)$$

де $\zeta_i = \zeta_i(x)$ є поліномами *i*-го степеня за *x*, які задовольняють умови ортогональності:

$$\begin{cases} \sum_{j=1}^{n} \zeta_{ij} = 0, \ \ i = 1, 2, ..., q ; \\ \sum_{j=1}^{n} \zeta_{ij} \zeta_{i'j} = 0, \ \ i \neq i' \end{cases}, \quad (7.33)$$

ζ_{ij} є *i*-м поліномом для точки x_j.

Квадратична форма, що мінімізується у методі найменших квадратів має вигляд:

$$Q = \sum_{j=1}^{n} \left(Y_j - \alpha'_0 - \alpha'_1 \zeta_{1j} - ... - \alpha'_q \zeta_{qj} \right)^2 . \quad (7.34)$$

Значення, що відповідають мінімуму (7.34), можна знайти, розв'язавши систему:

$$\begin{cases} \dfrac{\partial Q}{\partial \alpha'_0} = -2 \sum_{j=1}^{n} \left(Y_j - \alpha'^*_0 - \alpha'^*_1 \zeta_{1j} - ... - \alpha'^*_q \zeta_{qj} \right) = 0 ; \\ \dfrac{\partial Q}{\partial \alpha'_i} = -2 (\sum_{j=1}^{n} Y_j \zeta_{ij} - \alpha'^*_0 \sum_{j=1}^{n} \zeta_{ij} - \alpha'^*_1 \sum_{j=1}^{n} \zeta_{1j} \zeta_{ij} - ... - \alpha'^*_i \sum_{j=1}^{n} \zeta_{ij}^2 - ... - \\ \qquad - \alpha'^*_q \sum_{j=1}^{n} \zeta_{qj} \zeta_{ij}) = 0, \ \ i = 1, 2, ..., q \end{cases} . \quad (7.35)$$

Звідси, використовуючи умови ортогональності (7.33), одержуємо:

$$\alpha'^{*}_{0} = \overline{Y};\ \ \alpha'^{*}_{i} = \frac{\sum_{j=1}^{n} Y_j \zeta_{ij}}{\sum_{j=1}^{n} \zeta_{ij}^2}\ . \tag{7.36}$$

Використовуючи умови ортогональності, можна одержати явний вигляд поліномів для випадку, коли значення x змінюються з рівним кроком ω:

$$\begin{cases} \zeta_{0j} = 1; \\ \zeta_{1j} = \nu_j - \overline{\nu_j}; \\ \zeta_{2j} = \zeta_{1j}^2 - \left(n^2 - 1\right)/12\ , \end{cases} \tag{7.37}$$

де $\nu_j = \left(x_{j+1} - x_1\right)/\omega.$

Поліноми вищих степенів одержують за рекурентною формулою:

$$\zeta_{r+1,j} = \zeta_{1j}\zeta_{rj} - \frac{r^2\left(n^2 - r^2\right)}{4\left(4r^2 - 1\right)}\zeta_{r-1,j}\,. \tag{7.38}$$

Розглянемо такий приклад. У табл. 7.3 наведено результати вимірювання деякої величини.

Таблиця 7.3

Емпіричні дані для побудови поліноміальної регресійної моделі

j	1	2	3	4	5	6	7	8	9	10
x_j	0	10	20	30	40	50	60	70	80	90
Y_j	23	29	41	60	79	88	83	61	33	27

Побудуємо модель досліджуваної залежності у вигляді полінома 5-го степеня:

$$z = \alpha'_0 + \alpha'_1\zeta_1 + \alpha'_2\zeta_2 + \alpha'_3\zeta_3 + \alpha'_4\zeta_4 + \alpha'_5\zeta_5,$$

де $\zeta_i = \sum_{t=1}^{i} \beta_{it} x^t$ – поліноми i-го степеня, які задовольняють умови ортогональності.

Дані, необхідні для розрахунку значень ζ_{ij} і коефіцієнтів α_i, наведено в табл. 7.4.

Легко перевірити, що для одержаних даних виконуються умови ортогональності, тобто суми значень ζ_{ij} $(i \neq 0)$ у кожному рядку й суми за і добутків вигляду $\zeta_{ij}\zeta_{kj}$ $(i \neq k)$ дорівнюють нулю.

За даними таблиці розраховуємо коефіцієнти α'_i (табл. 7.5).

Таблиця 7.4

Результати розрахунку допоміжних параметрів для поліноміальної регресійної моделі та оцінок значень досліджуваної величини

v_j	1	2	3	4	5	6	7	8	9	10
ζ_{0j}	1	1	1	1	1	1	1	1	1	1
ζ_{1j}	−4,5	−3,5	−2,5	−1,5	−0,5	0,5	1,5	2,5	3,5	4,5
ζ_{2j}	12	4	−2	−6	−8	−8	−6	−2	4	12
ζ_{3j}	−25,2	8,4	21	18,6	7,2	−7,2	−18,6	−21	−8,4	25,2
ζ_{4j}	43,2	−52,8	−40,8	7,2	43,2	43,2	7,2	−40,8	−52,8	43,2
ζ_{5j}	−60	140	−10	−110	−60	60	110	10	−140	60
$y_j\zeta_{0j}$	23	29	41	60	79	88	83	61	33	27
$y_j\zeta_{1j}$	−103,5	−101,5	−102,5	−90	−39,5	44	124,5	152,5	115,5	121,5
$y_j\zeta_{2j}$	276	116	−82	−360	−632	−704	−498	−122	132	324
$y_j\zeta_{3j}$	−579,6	243,6	861	1116	568,8	−633,6	−1544	−1281	−277,2	680,4
$y_j\zeta_{4j}$	993,6	−1531	−1673	432	3413	3802	597,6	−2489	−1742	1166
$y_j\zeta_{5j}$	−1380	4060	−410	−6600	−4740	5280	9130	610	−4620	1620
Y_i^*	24,09	30,1	42,23	60,96	79,8	89,74	83,86	61,81	34,38	28,03

Таблиця 7.5

Результати розрахунку параметрів побудованої регресійної моделі

i	0	1	2	3	4	5
$\sum_{j=1}^{n}\zeta_{ij}^2$	10	82,5	528	3089	16474	78000
$\sum_{j=1}^{n}y_j\zeta_{ij}$	524	121	−1550	−845,4	2969	2950
α'_i	52,4	1,467	−2,936	−0,2737	0,1802	0,03782
$\sigma^2_{a'_j}$	0,038	0,0036	0,00072	0,00012	0,000023	0,000005
$\sigma_{a'_j}$	0,20	0,068	0,027	0,011	0,0048	0,0022

Оцінками дисперсії цих коефіцієнтів є величини $\sigma^2_{\alpha'_i} = \dfrac{\sum_{j=1}^{n}\left(y_i - y_i^*\right)^2}{(n-q-1)\sum_{j=1}^{n}\zeta_{ij}^2}$.

Звідси за формулами (7.32, 7.38) одержуємо оцінки Y^*_i значень досліджуваної величини Y у точках $x = x_j$, які наведені у табл. 7.3 і є досить близькими до її емпіричних значень.

На рис. 7.3 наведено графіки вихідних даних та побудованої регресійної моделі.

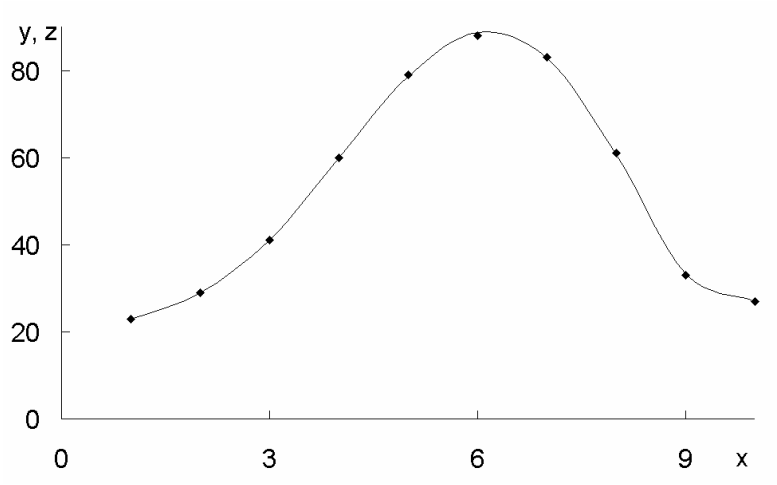


Рис. 7.3. Графіки вихідних даних (крапки) й побудованої поліноміальної регресійної моделі (лінія)

7.4. Однофакторні моделі інших типів

Апроксимацію емпіричних залежностей тригонометричними багаточленами називають **гармонічним аналізом**. У цьому випадку модель має вигляд:

$$z(x) = \alpha_0 + \sum_{k=1}^{r} \alpha_k \cos\frac{2\pi kx}{T} + \sum_{k=1}^{r} \beta_k \sin\frac{2\pi kx}{T}, \qquad (7.39)$$

де T – період спостереження апроксимованої залежності;
r – кількість гармонік ($r < n/2$);
n – кількість частин, на які розділений період T.

Її параметри визначають за формулами:

$$\alpha_0 = \frac{1}{n}\sum_{k=0}^{n} y_k;$$

$$\alpha_m = \frac{2}{n}\sum_{k=0}^{n} y_k \cos\frac{2\pi km}{n}, \quad m = 1, 2, ..., r; \qquad (7.40)$$

$$\beta_m = \frac{2}{n}\sum_{k=0}^{n} y_k \sin\frac{2\pi km}{n}, \quad m = 1, 2, ..., r,$$

де y_k – значення апроксимованої функції у точках $x_k = \frac{kT}{n}$.

На рис. 7.4 наведено графіки емпіричних даних і відповідних регресійних моделей, побудованих у вигляді тригонометричних рядів, для m рівних 2, 3, 4 й 5. Видно, що зі збільшенням кількості членів тригонометричного ряду різниця між моделлю та емпіричними точками зменшується. Добре видно також, що найбільшу похибку модель дає поблизу меж відрізка, на якому визначені емпіричні дані.

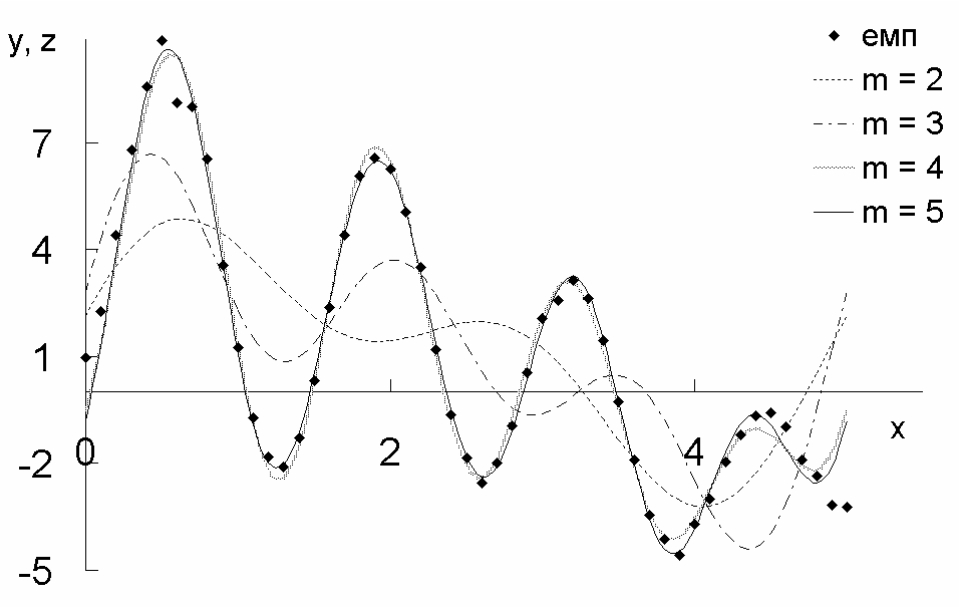


Рис. 7.4. Графіки вихідних даних і побудованої тригонометричної регресійної моделі

Моделі, що мають вигляд **модифікованої показникової функції**, записують як $z(x) = a + bc^x$ або $z(x) = a + bc^{-x}$. Якщо емпіричні точки є рівновіддаленими одна від одної, тобто крок емпіричної залежності за x є сталим, то можна перейти від незалежної змінної x до порядкового номера відповідної точки i та розрахувати параметри моделі першого типу за формулами:

$$c = \frac{(N-1)\sum_{i=1}^{N-1} y_i y_{i+1} - \sum_{i=1}^{N-1} y_i \sum_{i=1}^{N-1} y_{i+1}}{(N-1)\sum_{i=1}^{N-1} y_i^2 - \left(\sum_{i=1}^{N-1} y_i\right)^2};$$

$$b = \frac{N\sum_{i=1}^{N} c^i y_i - \sum_{i=1}^{N} c^i \sum_{i=1}^{N} y_i}{N\sum_{i=1}^{N} c^{2i} - \left(\sum_{i=1}^{N} c^i\right)^2}; \quad a = \frac{\sum_{i=1}^{N} y_i - b\sum_{i=1}^{N} c^i}{N}. \qquad (7.41)$$

У моделі другого типу параметри необхідно визначати за однією з ітераційних процедур мінімізації функціонала, що характеризує суму квадратів відхилень рівнів емпіричних точок від моделі. Зокрема у цьому випадку можна використовувати метод деформівного багатогранника.

Крива Гомперця описується рівняннями $\hat{y}_t = ab^{c^t}$ або $\hat{y}_t = ab^{c^{-t}}$, які логарифмуванням зводяться до узагальненої показникової функції першого або другого типу, відповідно.

Логістична функція $\hat{y}_t = \dfrac{1}{a + bc^t}$ або $\hat{y}_t = \dfrac{1}{a + bc^{-t}}$ зводиться до модифікованої показникової перетворенням $y^* = 1/\hat{y}_t = a + bc^{\pm t}$.

Модифіковану показникову функцію використовують як модель у випадках, коли досліджуваній залежності властиве насичення, тобто при збільшенні значень незалежної змінної відгук поступово наближається до певного граничного значення, а його прирости наближуються до нуля. У таких випадках існує певний обмежувальний фактор, вплив якого збільшується із зростанням досягнутого рівня. Значення рівня насичення, як правило, можна задати, виходячи з наявних даних про об'єкт дослідження. У такому разі інші параметри моделі можна визначити методом найменших квадратів після її лінеаризації.

Якщо вплив обмежувального фактора виявляється лише після досягнення певного рівня розвитку процесу, слід використовувати **моделі *S*-подібного зростання**, до яких належать крива Гомперця і логістична функція. Вони описують процеси, в яких темп зростання поступово збільшується на початкових стадіях і поступово зменшується в кінці. При цьому слід ураховувати, що крива Гомперця є асиметричною, а логістична крива симетрична стосовно точки перегину. Процес побудови й дослідження логістичної моделі називають логістичним аналізом. Логістичну криву часто називають законом зростання, оскільки вона описує залежність кількості популяції або її біомаси від часу.

При побудові регресійних моделей загального вигляду часто використовують методи, які базуються на мінімізації функціоналів виду (7.1). Для функціонала (7.1а) це зумовлює необхідність розв'язання системи:

$$\left\{ \frac{\partial}{\partial \alpha_k} \sum_{i=1}^{n} \left[z(x_i) - y_i \right]^2 = 0. \right. \qquad (7.42)$$

Якщо вона є системою лінійних рівнянь, застосовують звичайні алгоритми розв'язання таких систем – Гауса, простих ітерацій, Зейделя тощо. В окремих випадках, зокрема при логістичному аналізі, розробляють спеціальні алгоритми. Якщо ж система (7.42) є нелінійною, використовують алгоритми нелінійної оптимізації: Ньютона – Рафсона, квазіньютонівські, спряжених градієнтів тощо.

7.5. Лінійні багатофакторні моделі

Як було зазначено вище, лінійну як за параметрами, так і за незалежними змінними регресійну модель можна записати у вигляді:

$$Y = \alpha_0 + \sum_{i=1}^{p} \alpha_j x_j + \varepsilon = X\alpha + \varepsilon, \qquad (7.43)$$

де Y – вектор-стовпчик відгуків, який має розмірність n ($n > p$);

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$ – матриця значень p незалежних змінних при n вимірюваннях;

α – вектор-стовпчик невідомих параметрів моделі, що має розмірність $p + 1$;

ε – вектор-стовпчик похибок моделі, який має розмірність n.

Оцінки параметрів моделі у методі найменших квадратів отримують мінімізацією скалярного добутку:

$$Q = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\alpha})^{T} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\alpha}), \qquad (7.44)$$

де символ "Т" позначає транспонування.

Система (7.42) у цьому випадку набуває вигляду:

$$-2\boldsymbol{X}^{T} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\alpha}) = 0.$$

Звідси маємо:

$$\boldsymbol{X}^{T}\boldsymbol{Y} = \boldsymbol{X}^{T}\boldsymbol{X}\boldsymbol{\alpha}.$$

Помножуючи обидві частини цієї рівності ліворуч на матрицю $(\boldsymbol{X}^{T}\boldsymbol{X})^{-1}$, одержимо:

$$(\boldsymbol{X}^{T}\boldsymbol{X})^{-1}(\boldsymbol{X}^{T}\boldsymbol{Y}) = (\boldsymbol{X}^{T}\boldsymbol{X})^{-1}(\boldsymbol{X}^{T}\boldsymbol{X})\boldsymbol{\alpha}.$$

Добуток $(\boldsymbol{X}^{T}\boldsymbol{X})^{-1}(\boldsymbol{X}^{T}\boldsymbol{X}) = \boldsymbol{E}$, де $\boldsymbol{E}$ – одинична матриця розмірності $p + 1$. Тому остаточно маємо:

$$\boldsymbol{\alpha} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}. \qquad (7.45)$$

Для лінійної моделі (7.45) є незміщеною оцінкою з найменшою дисперсією вектора a.

Коваріаційною матрицею вектора a є:

$$\Sigma = \sigma^{2}(X^{T}X)^{-1}, \qquad (7.46)$$

де σ^{2} – дисперсія похибки.

Елементами головної діагоналі коваріаційної матриці є дисперсії компонентів вектора a, а позадіагональні компоненти є значеннями відповідних коефіцієнтів коваріації.

Для перевірки значущості регресії використовують *F*-критерій Фішера, розрахункове значення якого обчислюють за формулою:

$$F = \frac{Q_R / (p+1)}{Q / (n-p-1)}, \tag{7.47}$$

де $Q_R = (X\alpha)^T (X\alpha)$ – сума квадратів відхилень, зумовлена регресією;

Q – сума квадратів відхилень спостережень від регресії, що визначається за формулою (7.44).

За умови виконання нульової гіпотези $H_0: \alpha = 0$ $F < F_{кр}$, де $F_{кр}$ – критичне значення статистики Фішера для заданого рівня значущості й кількостей степенів вільності $(p+1)$ та $(n-p-1)$.

Значущість окремих коефіцієнтів регресії перевіряють за допомогою критерію:

$$t_j = \frac{\alpha_j}{\hat{s}\sqrt{(X^T X)^{-1}_{jj}}}, \tag{7.48}$$

де $\hat{s} = \sqrt{\frac{1}{n-p-1}Q}$ – незміщена оцінка стандартного відхилення залишків моделі. За умови виконання нульової гіпотези $H_0: \alpha_j = 0$ статистика критерію підпорядковується *t*-розподілу Стьюдента з кількістю степенів вільності $n - p - 1$.

Інтервальною оцінкою для коефіцієнта α_j є:

$$\alpha_j \in \left[\hat{\alpha}_j - t\hat{s}\sqrt{(X^T X)^{-1}_{jj}}; \hat{\alpha}_j + t\hat{s}\sqrt{(X^T X)^{-1}_{jj}} \right]. \tag{7.49}$$

Одним з ускладнень, що можуть виникати при побудові регресійних моделей за наявності декількох предикторів, є можлива нерівноточність спостережень. У найпростішому випадку це можна врахувати за допомогою коваріаційної матриці такого вигляду:

$$\Omega = \begin{pmatrix} \sigma_1^2 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & \sigma_n^2 \end{pmatrix}, \tag{7.50}$$

де σ_i^2 – дисперсія *i*-го спостереження. У цьому випадку формула (7.45) набуває вигляду:

$$\alpha = \left(X^T \Omega^{-1} X \right)^{-1} X^T \Omega^{-1} Y. \tag{7.51}$$

Такий метод побудови регресійних моделей називають **зваженим методом найменших квадратів**. Вперше його було запропоновано К. Гаусом в 1809 р.

У цьому випадку коваріаційна матриця визначається за формулою:

$$\Sigma = \sigma^2 \left(X^T \Omega^{-1} X \right)^{-1}. \tag{7.52}$$

Для перевірки значущості регресії використовують *F*-критерій Фішера, розрахункове значення якого обчислюють за формулою (7.47), де $Q_R = \left(X\alpha \right)^{\mathrm{T}} \Omega^{-1} \left(X\alpha \right)$, $Q = \left(Y - X\alpha \right)^{\mathrm{T}} \Omega^{-1} \left(Y - X\alpha \right)$.

Значущість окремих коефіцієнтів регресії перевіряють за допомогою критерію:

$$t_j = \frac{\alpha_j}{\hat{s}^2_{\alpha_j}} = \frac{\alpha_j}{\hat{s}\sqrt{\left(X^T \Omega^{-1} X \right)^{-1}_{jj}}}, \tag{7.53}$$

де $\hat{s}^2_{\alpha_j}$ – *j*-й діагональний елемент коваріаційної матриці.

Інтервальною оцінкою для коефіцієнта α_j є:

$$\alpha_j \in \left[\hat{\alpha}_j - t\hat{s}_{\alpha_j}; \hat{\alpha}_j + t\hat{s}_{\alpha_j} \right]. \tag{7.54}$$

Іншим можливим ускладненням є наявність взаємозв'язку між предикторами. Припустимо, що існує лінійна залежність між компонентами вектора ***X***:

$$\nu_1 x_1 + \nu_2 x_2 + ... + \nu_m x_m = 0. \tag{7.55}$$

Що ближчою є ліва частина (7.55) до нульового вектора, то сильнішою є мультиколінеарність. Граничний випадок точного виконання рівності (7.55) називають **строгою мультиколінеарністю**. У цьому випадку визначник $|X^{\mathrm{T}}X| = 0$ й використовувати формулу (7.45) неможливо. Випадок $|X^{\mathrm{T}}X| \approx 0$ називають **мультиколінеарністю**.

Мультиколінеарність зумовлює нестійкість обчислювальної процедури через високу похибку обчислення оберненої матриці, тобто додавання нових даних призводить до істотної зміни оцінок параметрів. Коефіцієнти регресійної моделі у цьому випадку виявляються сильно корельованими один з одним, а довірчий рівень і дисперсія їх оцінок – підвищеними. Внаслідок цього інтерпретація результатів стає неможливою, а значення окремих коефіцієнтів – статистично незначущими.

Наявність мультиколінеарності можна перевірити шляхом дослідження кореляційної матриці ***R*** нормованих і центрованих вихідних даних [16]. У цьому випадку: $|R| << 1$ і для окремих елементів матриці $|r_{ij}| \geq 0{,}9$ при $i \neq j$.

Свідченням мультиколінеарності є також **погана зумовленість** матриці ($X^{\mathrm{T}}X$), яку визначають як відношення максимального власного числа матриці до мінімального. Якщо $\frac{\lambda_{\max}}{\lambda_{\min}} \geq 10^5$, то це є свідченням сильної мультиколінеарності вихідних даних.

Існує декілька способів корегування мультиколінеарності. Найпростішим є стандартизація й центрування даних.

Інший підхід передбачає залучення додаткової інформації, зокрема, збільшення обсягу вибірки, за якою оцінюють значення параметрів моделі. Проте, в більшості випадків це неможливо, особливо при дослідженні соціально-економічних систем.

Ще один підхід ґрунтується на зменшенні розмірності простору факторів. Найпростіше це можна зробити шляхом виявлення найсильніше корельованих змінних і об'єднання їх в один фактор. Але це дає позитивні результати лише за умови, що таке об'єднання є теоретично обґрунтованим.

В окремих випадках для усунення мультиколінеарності відкидають одну чи декілька сильно пов'язаних змінних, що призводить до появи нових похибок. У такому випадку необхідно визначити, яка з похибок є більш істотною. Для цього можна почергово виключати пов'язані змінні й порівнювати одержувані результати. Інший варіант цього підходу передбачає послідовне додавання нових факторів і перевірку того, покращує він модель чи ні.

Одним з методів відбору найбільш істотних факторів є процедура **покрокової регресії**. Вона може бути організована як у напрямі зменшення кількості факторів, що враховують у моделі, так і в зворотному. У першому випадку спочатку будують модель, що враховує всі фактори і перевіряють їх значущість. Для цього перевіряють нульові гіпотези $\alpha_j = 0$, використовуючи статистику:

$$F_j = \frac{\hat{\alpha}_j^{\,2}}{D\left[\hat{\alpha}_j\right]}, \tag{7.56}$$

де $D\left[\hat{\alpha}_i\right] = \dfrac{\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2}{n-p}\left(X^{-1}X\right)_{ii}^{-1}$.

За умови справедливості нульової гіпотези вона має розподіл Фішера з кількістю степенів вільності 1 та $(n - p)$. Найменше значення F_j порівнюють з граничною величиною F_0. Якщо $F_j < F_0$, то j-й фактор виключають із моделі й будують нову модель з меншою кількістю факторів. В іншому випадку модель залишають без змін.

У випадку, коли процедуру організовують у напрямі збільшення кількості факторів, на першому етапі визначають коефіцієнти кореляції між кожним фактором і відгуком, а потім будують однофакторну модель регресії, яка враховує лише фактор x_1, що має найбільший (за модулем) коефіцієнтом кореляції.

Якщо перевірка моделі встановлює значущість обраного фактора, то наступним кроком є визначення F-статистики (7.56) для факторів, що залишилися. До нової моделі включають фактор x_1, а також фактор x_m, для якого значення цієї статистики є найбільшим. Потім визначають значущість отриманої моделі й розраховують F-статистики F_1 та F_m. Меншу з них порівнюють з граничним значенням F_0 і за виконання умови $F_j < F_0$ відповідний фактор виключають із моделі.

Таку процедуру здійснюють на кожному наступному кроці. Побудову моделі закінчують, коли найбільше значення F-статистики для факторів, що не включені до неї, не перевищує граничного значення F_0. Процес також закінчують, якщо до моделі включено всі досліджувані фактори або якщо перевищено граничну кількість кроків.

Можна також перетворити множину вихідних факторів до меншої кількості нових взаємноортогональних факторів. У цьому випадку використовують метод головних компонент та інші процедури факторного аналізу.

Розглянуті вище методи усунення мультиколінеарності розраховані на отримання незміщених оцінок параметрів регресійних моделей. Альтернативою їм є **методи зміщеного оцінювання**, зокрема гребеневі оцінки, редуковані оцінки, оцінки Марквардта, оцінки Хоккінса тощо. Якщо коваріаційні матриці є погано зумовленими, при обчисленні оцінок коефіцієнтів регресії застосовують також метод регуляризації О.М. Тихонова. Завданням оцінювання при їх застосуванні є одержання значень параметрів регресії, які були б стійкими навіть за умови сильної спряженості незалежних змінних.

У граничному випадку, коли матриця $X'X$ є одиничною, оцінки, одержувані за формулою (7.55), є незміщеними і мають мінімальну дисперсію. Із збільшенням власних чисел матриці $X'X$ відстань між оцінками $\hat{\beta}$ та істинними β збільшується. Крім того, стає можливою зміна напряму впливу вхідних змінних на вихідні.

На сьогодні розроблено понад 80 алгоритмів зміщеного оцінювання параметрів лінійних регресійних моделей. Їх поділяють на такі групи: методи звичайних гребеневих оцінок, методи узагальнених гребеневих оцінок, методи оцінок дробового рангу, методи стиснутих оцінок. Всі вони є лінійними перетвореннями оцінок, одержуваних МНК.

У **методах гребеневого аналізу** ставиться завдання отримання оцінок з мінімальною дисперсією. Оцінка гребеневої регресії α_k є лінійним перетворенням оцінки $\boldsymbol{\alpha}$, отриманої МНК, і залежить від параметра k і матриці вихідних даних $\boldsymbol{X}$. Її можна записати у вигляді:

$$\alpha_k = \left(X^T X - kE \right)^{-1} X^T Y . \qquad (7.57)$$

Ефективність методів гребеневих оцінок залежить від статистичних характеристик вихідної інформації й оптимального вибору параметра k. Метод гребеневого аналізу, як і інші методи зміщеного оцінювання параметрів регресійних моделей, за певних умов можуть бути обґрунтовані теоретично, але в більшості практичних ситуацій перевірити виконання цих умов неможливо. Тому застосування цих методів потребує певної обережності.

Параметри нелінійних регресійних моделей за наявності декількох незалежних змінних зазвичай оцінюють з використанням чисельних методів нелінійної мінімізації функціонала (7.1а). При цьому істотне значення має вибір початкового наближення. У багатьох випадках це завдання може бути істотно спрощено з використанням методів планування експерименту, які дають змогу визначити оптимальні для подальшого аналізу плани, тобто точки простору незалежних ознак, у яких потрібно здійснити вимірювання значень відгуків.

7.6. Інші типи багатофакторних моделей

Для статичних систем багатофакторні регресійні моделі зазвичай задають у вигляді полінома:

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{\substack{i=1 \\ i \neq j}}^{k} \beta_{ij} x_i x_j + \sum_{i=1}^{k} \beta_{ii} x_i^2 + ... \qquad (7.58)$$

При цьому найчастіше обмежуються поліномами другого степеня.

У деяких випадках, зокрема при проектуванні теплотехнічних і гідродинамічних систем, а також при дослідженні економічних систем модель задають у вигляді степеневої функції:

$$y = \gamma x_1^{\beta_1} x_2^{\beta_2} .... \qquad (7.59)$$

У такому випадку модель можна лінеаризувати логарифмуванням:

$$\ln y = \ln \gamma + \beta_1 \ln x_1 + \beta_2 \ln x_2 + ...$$

та переходом до нових змінних: $y' = \ln y$, $x'_1 = \ln x_1$, $x'_2 = \ln x_2$ … Після цього аналіз моделі здійснюють за допомогою методів, описаних у попередньому підрозділі.

Відомим прикладом моделі такого типу є відома в економічній теорії **функція Кобба – Дугласа**:

$$Q = Q_0 \left(\frac{L}{L_0} \right)^a \left(\frac{K}{K_0} \right)^b, \tag{7.60}$$

де Q – обсяг виробництва;

L – трудові ресурси;

K – капітал.

Індекс “0” відповідає певним фіксованим значенням цих параметрів. Логарифмуванням ця модель приводиться до лінійної:

$$\ln Q = \ln Q_0 + a \ln \left(\frac{L}{L_0} \right) + b \ln \left(\frac{K}{K_0} \right). \tag{7.61}$$

7.7. Перевірка адекватності регресійних моделей

Основні методи перевірки адекватності регресійних моделей ґрунтуються на таких трьох властивостях їх залишків.

По-перше, для адекватної моделі дисперсія залишків має бути близькою до дисперсії емпіричних точок. При цьому припускають, що дисперсії всіх емпіричних точок є однаковими. У випадку, коли для кожної точки здійснюють декілька вимірювань значення відгуку, останнє припущення можна перевірити за допомогою критеріїв Кокрена або Бартлетта.

Причиною неадекватності при невиконанні цієї властивості є використання надмірно спрощених або ускладнених регресійних моделей. Відомо, наприклад, що за наявності n емпіричних точок можна побудувати поліноміальну модель $n - 1$ порядку, яка пройде строго через всі ці точки. Але використовувати такий поліном як регресійну модель за наявності похибок емпіричних даних, очевидно, недоцільно. З іншого боку, якщо степінь полінома буде надто малим, то він не відтворюватиме істотних рис досліджуваної залежності, тому існує певне оптимальне значення степеня такого полінома. Як критерій виконання цієї властивості часто застосовують такий критерій:

$$\frac{S}{\Delta^2} \le F, \; \frac{\Delta^2}{S} \le F, \tag{7.62}$$

де S – значення цільового функціонала (7.1а);

Δ^2 – сума квадратів похибок емпіричних даних по всіх точках;

F – критичне значення критерію Фішера для вибраного рівня значущості й кількостей степенів вільності, що дорівнюють $n - 1$.

Невиконання першої умови свідчить про надмірну спрощеність моделі, зокрема про необхідність збільшення порядку поліноміальної моделі. Невиконання другої умови є свідченням того, що модель треба спростити,

наприклад зменшити порядок полінома. У деяких випадках друга умова може не виконуватися навіть для однофакторних лінійних моделей. Найчастіше це може бути наслідком свідомого підганяння емпіричних даних під заздалегідь задану модель. Це часто роблять у навчальних задачах, але на практиці такий результат свідчить про навмисне викривлення первинних даних. Іншою причиною може бути неправильна (завищена) оцінка похибки емпіричних даних. Це може бути пов'язано, зокрема, з нехтуванням зміною дисперсії емпіричних даних при їх попередній обробці.

Інша властивість залишків, яку перевіряють при визначенні адекватності моделі, полягає в тому, що вони мають підпорядковуватися нормальному закону розподілу з нульовим математичним сподіванням і однаковими дисперсіями. Перевірку цих властивостей можна здійснити за допомогою критеріїв, що описані у розділі 2.

На рис. 7.5 показано деякі типові випадки порушення вказаних властивостей, які можуть бути виявлені при візуальному аналізі ряду залишків.

У випадку а) на графіку є так звані викиди – точки, що аномально сильно відхиляються від середнього значення. У випадку б) дисперсія залишків помітно зменшується при зміщенні вправо. У випадку в) ряд залишків не є випадковим, що свідчить про наявність неврахованих істотних закономірностей у моделі.

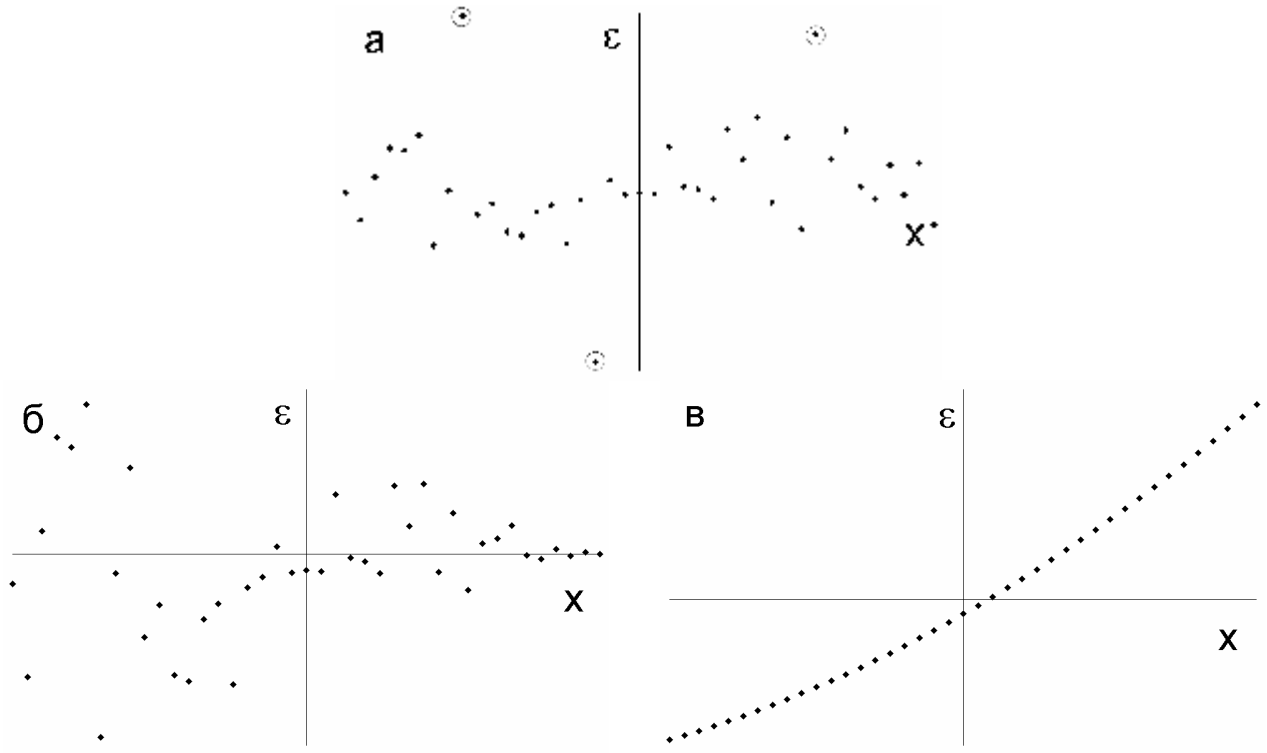


Рис. 7.5. Приклади порушення властивостей залишків неадекватних моделей

Третьою властивістю є те, що залишки адекватної регресійної моделі мають бути некорельованими випадковими величинами. Наявність автокореляції першого порядку перевіряють за допомогою **критерію Дарбіна − Уотсона**:

$$d = \frac{\sum_{i=2}^{n} \left(e_i - e_{i-1}\right)^2}{\sum_{i=1}^{n} e_i^2}, \qquad (7.63)$$

де n – кількість емпіричних точок. Для адекватної моделі має виконуватися умова $d \approx 2$. Близькі до нуля значення d свідчать про наявність додатної автокореляції, а значення, що наближаються до 4, – про наявність від'ємної автокореляції.

Цей критерій було розроблено в 1950 р. британським статистиком Джеймсом Дарбіном та австралійським статистиком Джеффрі Стюартом Уотсоном.

Наявність автокореляції вищих порядків перевіряють шляхом дослідження автокореляційної функції. Про наявність автокореляції в цьому випадку свідчить збільшення абсолютних значень коефіцієнта автокореляції при певних значеннях параметра зсуву. На рис 7.6 показано приклади автокореляційних функцій для ряду, що є білим шумом, (ліворуч) та рядом, який змінюється за синусоїдальним законом, (праворуч).

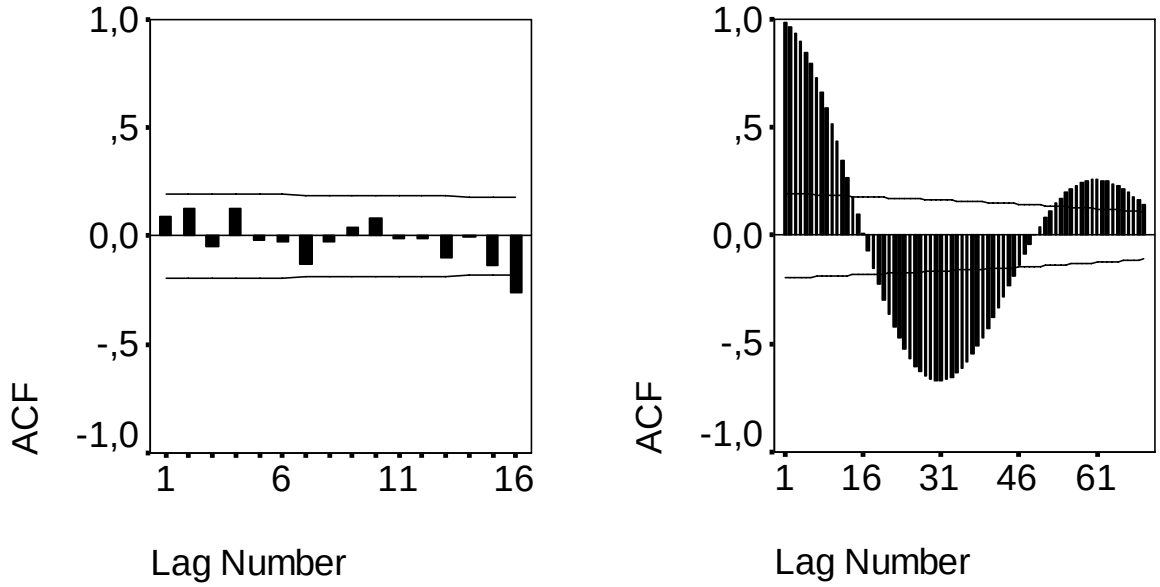


Рис. 7.6. Приклади автокореляційних функцій деяких рядів

Горизонтальними лініями на цих графіках показано довірчі інтервали для нульових значень коефіцієнта автокореляції. З наведених графіків добре видно, що у першому випадку автокореляція є практично відсутньою, а в другому – для певних значень параметра зсуву спостерігається істотна додатна або від'ємна автокореляція.

7.8. Побудова однофакторних регресійних моделей в електронних таблицях MS Excel

Найпростішим випадком є побудова одновимірної лінійної регресійної моделі. Її параметри можна визначити безпосередньо за формулами 7.5. Але в електронних таблицях MS Excel це можна зробити за допомогою вбудованих формул та пакета аналізу.

Розглянемо такий приклад. Нехай ми маємо дві пов'язані вибірки обсягом по 41 елементу. Елементами першої вибірки x_i є числа від –2 до 2, взяті у порядку зростання з кроком 0,1. Елементи другої вибірки розраховуємо за формулою $y_i = 2x_i + 1 + \varepsilon_i$, де ε_i – нормально розподілені випадкові числа з математичним сподіванням 0 та стандартним відхиленням 0,2.

Для побудови лінійної моделі можна скористатися функцією "ЛИНЕЙ()". На рис. 7.7 показано діалогове вікно задання її параметрів.

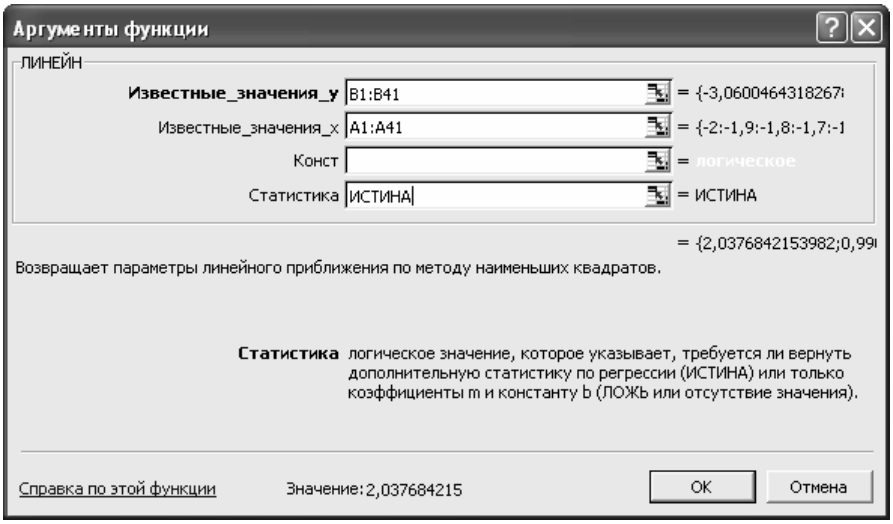
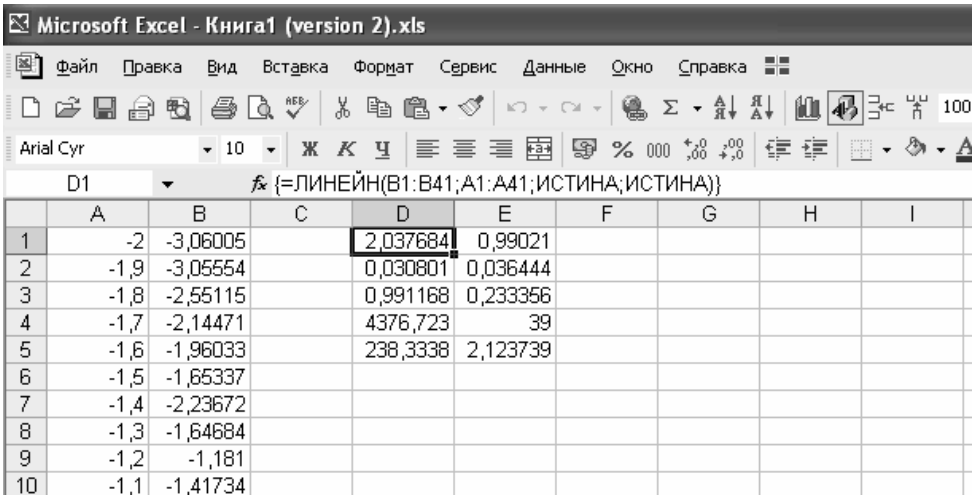


Рис. 7.7. Діалогове вікно задання параметрів функції "ЛИНЕЙ()"

У перших двох комірках задаємо посилання на комірки, що містять значення змінних *Y* та *X*, відповідно. У третій комірці вказуємо необхідність підбору константи (якщо значення дорівнює "ИСТИНА" або відсутнє, константу необхідно обчислити; якщо значення дорівнює "ЛОЖЬ", то константа береться рівною нулю). В останній комірці вказуємо необхідність виведення підсумкової статистики ("ИСТИНА") чи тільки коефіцієнтів моделі ("ЛОЖЬ").

Уведення формули необхідно здійснювати як формулу масиву. Для цього потрібно виділити на робочому аркуші масив суміжних комірок об-

сягом 2×5, записати формулу й після цього натиснути клавішу F2, а потім одночасно клавіші Ctrl+Shift+Enter. При цьому отримуємо масив результатів, наведений на рис. 7.8.



	A	B	C	D	E	F	G	H	I
1	-2	-3,06005		2,037684	0,99021				
2	-1,9	-3,05554		0,030801	0,036444				
3	-1,8	-2,55115		0,991168	0,233356				
4	-1,7	-2,14471		4376,723	39				
5	-1,6	-1,96033		238,3338	2,123739				
6	-1,5	-1,65337							
7	-1,4	-2,23672							
8	-1,3	-1,64684							
9	-1,2	-1,181							
10	-1,1	-1,41734							

Рис. 7.8. Результати обчислення параметрів лінійної регресії

Результати наведено у комірках D1:E5. При цьому: D1, E1 – це значення коефіцієнтів рівняння регресії $y = ax + b$; D2, E2 – стандартні відхилення коефіцієнтів моделі a й b; D3 – коефіцієнт детермінації моделі; E3 – стандартне відхилення для значень y; D4 – F-статистика; E4 – кількість степенів вільності для F-статистики; D5 – регресійна сума квадратів; E5 – сума квадратів залишків.

Крім лінійної, в електронних таблицях MS Excel можна побудувати степеневу регресійну модель вигляду $y = ab^x$, використовуючи вбудовану формулу “=ЛГРФПРИБЛ()”. Застосування цієї формули є таким самим, як і формули “ЛИНЕЙ()”.

Іншим варіантом побудови регресійної моделі в електронних таблицях MS Excel є застосування пакету аналізу. Для цього обираємо у головному меню: Сервіс/Аналіз даних/Регресія. Після цього відкривається діалогове вікно (рис. 7.9).

У цьому вікні позначаємо посилання на комірки, де містяться значення змінних x, y. Позначку “Метки” робимо у випадку, коли перший стовпчик або перший рядок вхідних даних містять заголовки. Позначку “Константа – ноль” робимо у випадку, коли вільний член моделі дорівнює нулю. У комірці “Уровень надежності” задаємо довірчий рівень (за умовчанням він дорівнює 0,95).

Далі позначаємо, куди саме слід виводити результати, а також необхідність виводу статистичних характеристик моделі та побудови графіку нормальної імовірності. Результати для тих самих вихідних даних, що і у попередньому випадку, показано на рис. 7.10.

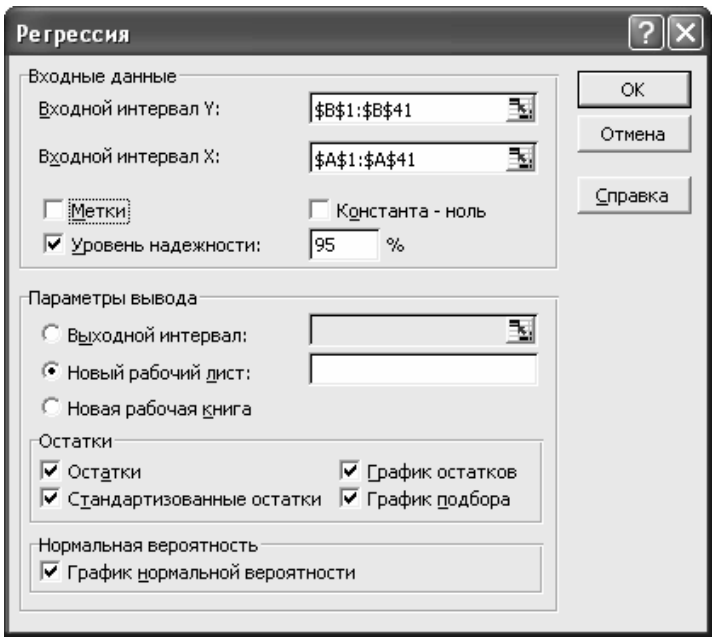


Рис. 7.9. Діалогове вікно побудови регресійної моделі в пакеті аналізу

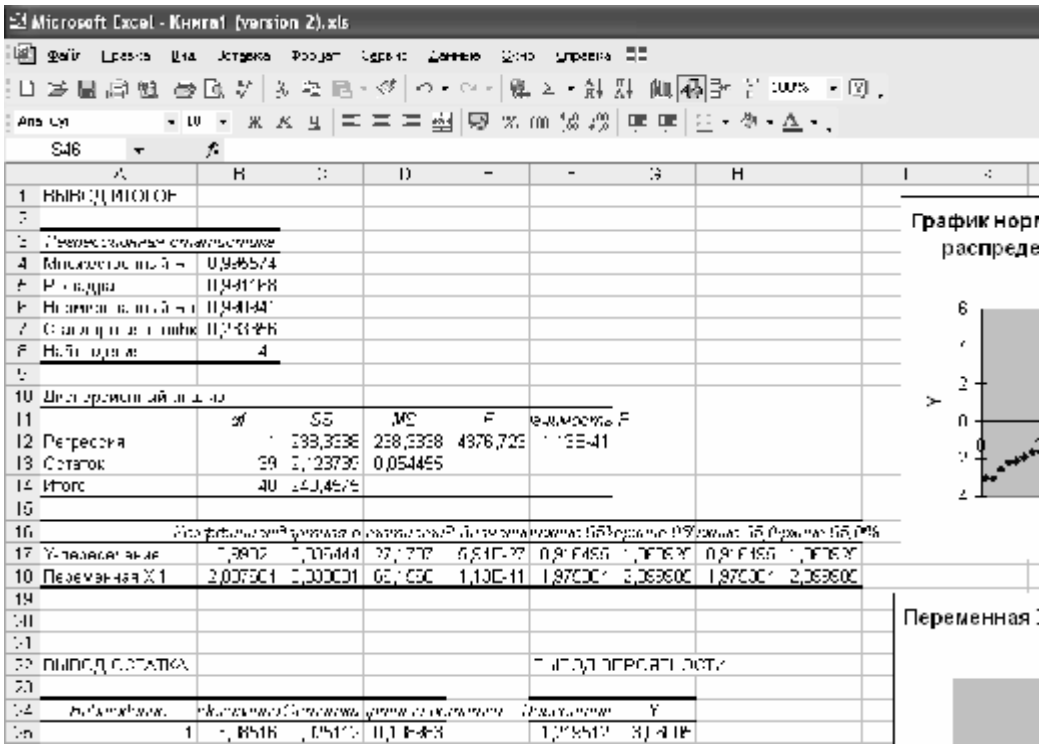


Рис. 7.10. Результати підбору параметрів регресійної моделі

Бачимо, що основні параметри є такими самими, що й у попередньому випадку. Проте слід зазначити, що за допомогою пакету аналізу ми можемо отримати більш докладну інформацію про властивості моделі. Крім того, використання пакету аналізу є простішим. Зокрема, воно не передбачає необхідності заздалегідь розраховувати й виокремлювати на робочому аркуші комірки для виводу результатів.

7.9. Побудова однофакторних регресійних моделей в пакеті SPSS

У пакеті SPSS також є різні засоби побудови регресійних моделей.

Для побудови лінійної моделі заносимо дані аркуш даних (рис. 7.11) й використовуємо пункти меню: Analyze/Regression/Linear.

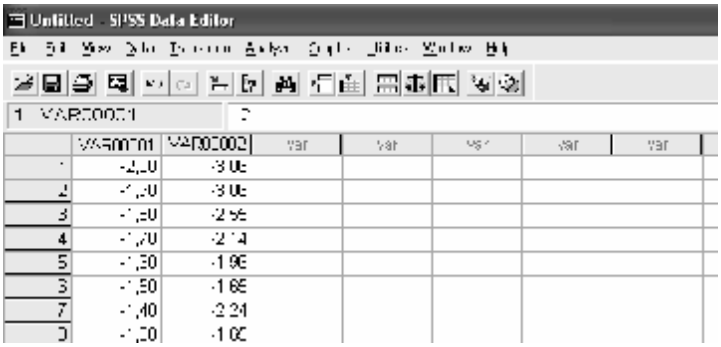


Рис. 7.11. Аркуш даних пакету SPSS при побудові регресійної моделі

При цьому відкривається діалогове вікно задання параметрів процедури (рис. 7.12). У цьому вікні необхідно вказати залежну й незалежну (або декілька незалежних) змінну. У випадку декількох незалежних змінних можна згрупувати їх у блоки й обрати методи вводу для різних блоків. Також при застосуванні зваженого методу найменших квадратів можна задати вагові коефіцієнти для незалежних змінних.

У діалоговому вікні “Statistics” (рис. 7.13) задаємо, які статистичні характеристики моделі необхідно показати у вікні результатів. Вивід коефіцієнтів кореляції між змінними й діагностика мультиколінеарності для одно факторної моделі не потрібні.

У вікні “Plots” (рис. 7.14) зазначаємо, які графіки необхідно вивести у вікні результатів.

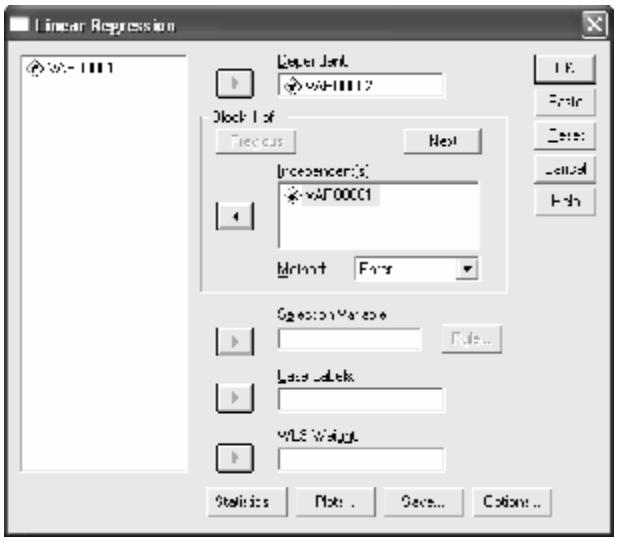


Рис. 7.12. Діалогове вікно задання параметрів побудови лінійної регресійної моделі в пакеті SPSS

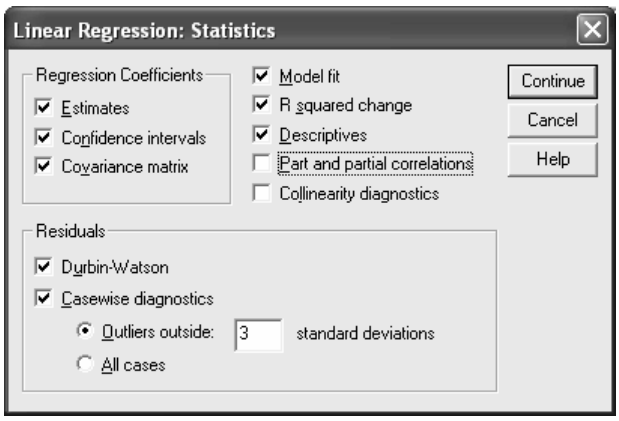


Рис. 7.13. Діалогове вікно задання параметрів статистики лінійної регресійної моделі

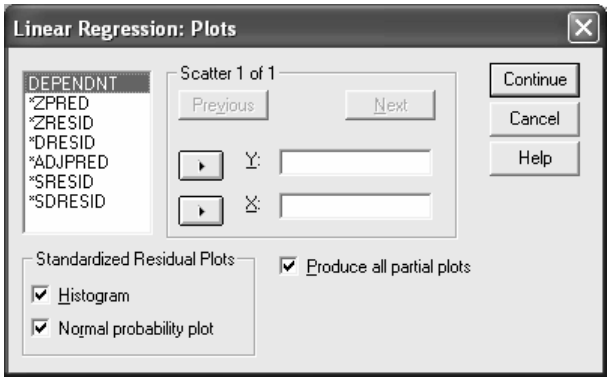


Рис. 7.14. Діалогове вікно задання графіків, які потрібно показати у вікні результатів

У діалоговому вікні “Save” (рис. 7.15) вказуємо, які результати необхідно зберегти у вікні даних як нові змінні. Зокрема можна задати формування таких змінних як передбачувані значення залежної змінної, залишки моделі тощо. Частина змінних потрібна тільки при побудові багатофакторних лінійних моделей.

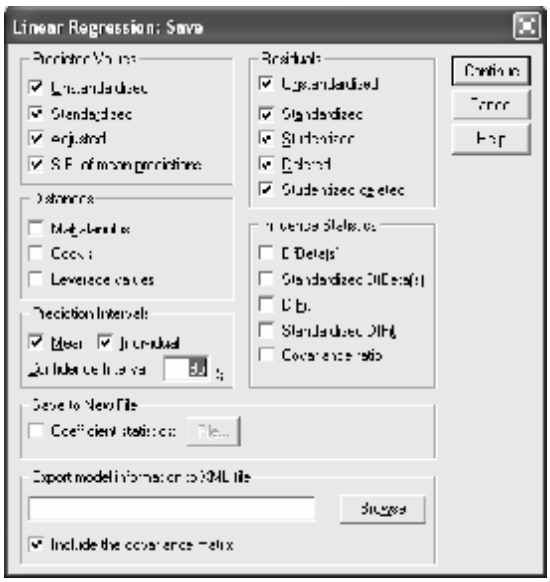


Рис. 7.15. Діалогове вікно задання нових змінних

У діалоговому вікні “Options” (рис. 7.16) задаємо додаткові параметри побудови моделі.

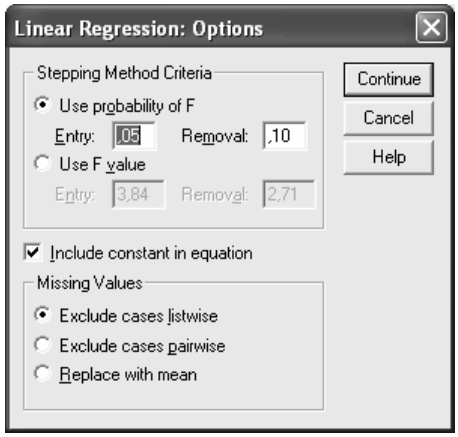


Рис. 7.16. Діалогове вікно задання додаткових параметрів побудови моделі.

Деякі результати наведено на рис. 7.17, 7.18. Бачимо, що вони збігаються з результатами, отриманими в електронних таблицях MS Excel, а також з вихідними даними. Але слід зазначити, що у пакеті SPSS ми маємо можливість отримати значно більше статистичних даних стосовно якості побудованої моделі.

Descriptive Statistics

	Mean	Std. Deviation	N
VAR00002	,9902	2,45182	41
VAR00001	,0000	1,19791	41

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics		Durbin-Watson
					R Square Change	F Change	
1	,996[a]	,991	,991	,23336	,991	4376,723	1,805

a. Predictors: (Constant), VAR00001

b. Dependent Variable: VAR00002

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	238,334	1	238,334	4376,723	,000[a]
	Residual	2,124	39	,054		
	Total	240,458	40			

a. Predictors: (Constant), VAR00001

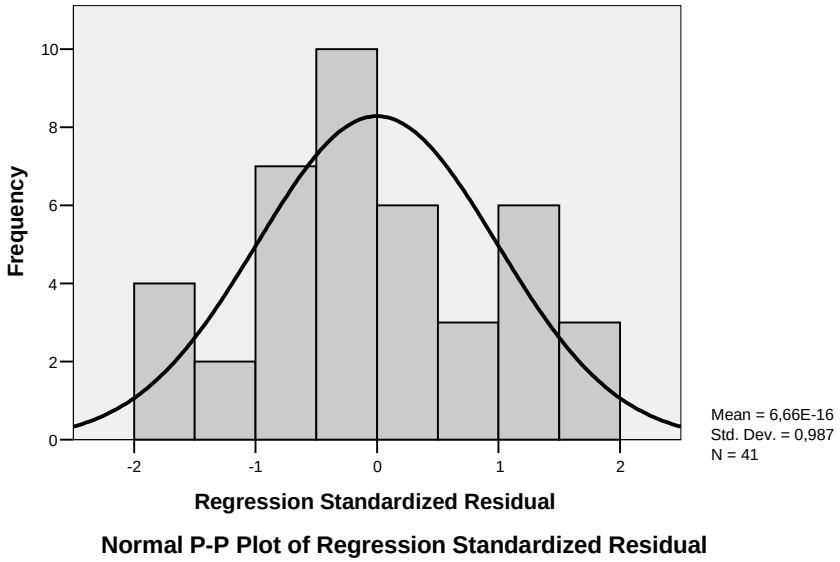
b. Dependent Variable: VAR00002

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-3,0852	5,0656	,9902	2,44097	41
Std. Predicted Value	-1,670	1,670	,000	1,000	41
Standard Error of Predicted Value	,036	,072	,050	,011	41
Adjusted Predicted Value	-3,0878	5,1024	,9899	2,44293	41
Residual	-,39495	,45477	,00000	,23042	41
Std. Residual	-1,692	1,949	,000	,987	41
Stud. Residual	-1,717	1,984	,001	1,013	41
Deleted Residual	-,40664	,47153	,00029	,24275	41
Stud. Deleted Residual	-1,763	2,066	,004	1,031	41
Mahal. Distance	,000	2,787	,976	,883	41
Cook's Distance	,000	,132	,027	,034	41
Centered Leverage Value	,000	,070	,024	,022	41

a. Dependent Variable: VAR00002

Рис. 7.17. Деякі результати побудови лінійної регресійної моделі у пакеті SPSS



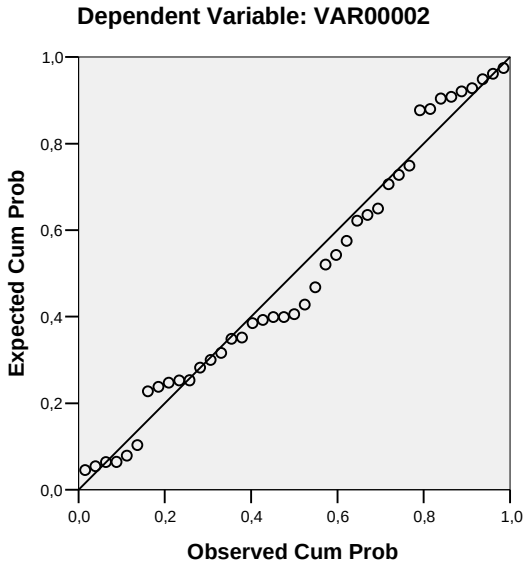


Рис. 7.18. Деякі графіки результатів побудови лінійної регресійної моделі

Розглянемо інший засіб побудови регресійних моделей у пакеті SPSS. Для цього внесемо на аркуш даних такий масив: незалежна змінна X є набором чисел від –2 до 2 з кроком 0,1; значення залежної змінної розраховано за формулою $yi = (2x^3_i - 3x^2_i + 5x_i + 2)\varepsilon_i$, де ε_i – елементим нормально розподіленої випадкової послідовності з математичним сподіванням 1 й стандартним відхиленням 0,2. Оберемо у головному меню: Analyze/Regression/Curve estimation. При цьому відкривається діалогове вікно задання параметрів процедури оцінювання параметрів кривих (рис. 7.19).

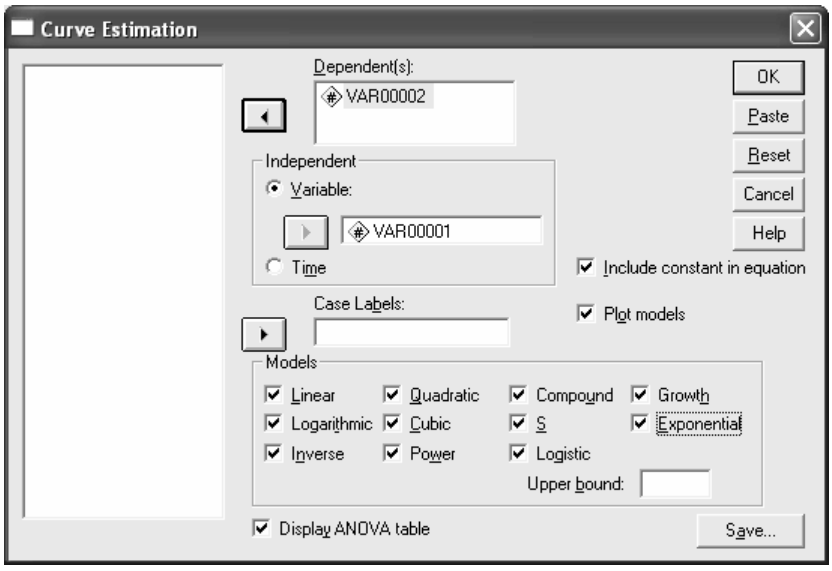


Рис. 7.19. Діалогове вікно задання параметрів процедури оцінювання кривої

У цьому вікні ми вказуємо залежну й незалежну змінні, необхідність включення вільного члена, побудови графіків моделей і таблиці ANOVA, а також типи оцінюваних моделей. Позначку “Case Labels” використовують для вибору типів маркерів при побудові графіків у випадку декількох залежних змінних. Кнопку “Save” використовують для формування нових змінних (передбачувані значення залежної змінної, залишки моделі, довірчі інтервали) на аркуші даних.

У випадку, що розглядається, не всі типи моделей можуть бути побудовані. Про це на аркуші результатів виводиться відповідне повідомлення. Зокрема обернену й *S*-подібну моделі неможливо побудувати через наявність нульового значення незалежної змінної; логарифмічну та степеневу – через наявність від’ємних значень незалежної змінної; складену, степеневу, *S*-подібну, зростання, експоненціальну й логістичну – через наявність від’ємних значень залежної змінної.

Тому залишаються три типи доступних моделей – лінійна, квадратична й кубічна. На рис. 7.20 побудовано їх графіки, з яких видно, що найбільш придатною є кубічна модель. Це відповідає вихідним даним. Основні результати для кубічної моделі наведено на рис. 7.21.

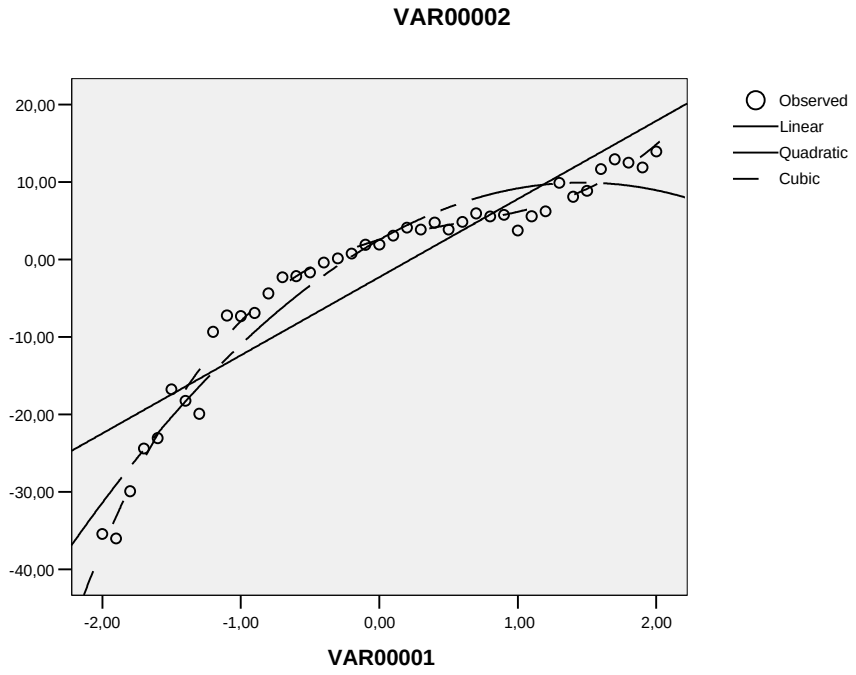


Рис. 7.20. Графіки побудованих моделей

Бачимо, що отримані результати задовільно збігаються з вихідними даними, але є помітні розбіжності у значеннях окремих коефіцієнтів моделі. Це може бути пов’язано з малим обсягом вихідної вибірки й високим рівнем стандартного відхилення при розрахунку вихідних значень залежної змінної.

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
,993	,986	,985	1,625

The independent variable is VAR00001.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	6874,087	3	2291,362	867,333	,000
Residual	97,748	37	2,642		
Total	6971,835	40			

The independent variable is VAR00001.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
VAR00001	5,063	,537	,459	9,420	,000
VAR00001 ** 2	-3,444	,203	-,330	-16,972	,000
VAR00001 ** 3	1,995	,196	,497	10,192	,000
(Constant)	2,540	,381		6,668	,000

Рис. 7.21. Основні результати підбору кубічної моделі

7.10. Побудова однофакторних регресійних моделей в пакеті MathCad

Для побудови лінійної моделі створимо масив даних, що містить 11 точок. Значення змінної *x* сформуємо у вигляді арифметичної прогресії з першим членом –5 та різницею 1. Значення змінної у сформуємо за формулою:

$$y = 2x + 1 + \varepsilon,$$

де ε – рівномірно розподілена випадкова величина, задана на відрізку [–2; 2].

На рис. 7.22, 7.23 показано робочі вікна з двома варіантами програми побудови моделі, а на рис. 7.24 – результат побудови моделі (однаковий для обох випадків).

$$x := (-5 \ -4 \ -3 \ -2 \ -1 \ 0 \ 1 \ 2 \ 3 \ 4 \ 5)^T$$

$$y := (-9.5 \ -8.6 \ -4.6 \ -3.4 \ 0.5 \ 2.8 \ 1.1 \ 4.6 \ 8.8 \ 7.4 \ 9.9)^T$$

$$line(x, y) = \begin{pmatrix} 0.818 \\ 1.98 \end{pmatrix}$$

$$f(t) := line(x, y)_0 + line(x, y)_1 \cdot x$$

Рис. 7.22. Фрагмент програми побудови лінійної однофакторної моделі за допомогою функції line(*x*, *y*)

$$x := (-5 \ -4 \ -3 \ -2 \ -1 \ 0 \ 1 \ 2 \ 3 \ 4 \ 5)^T$$

$$y := (-9.5 \ -8.6 \ -4.6 \ -3.4 \ 0.5 \ 2.8 \ 1.1 \ 4.6 \ 8.8 \ 7.4 \ 9.9)^T$$

$$intercept(x, y) = 0.818$$

$$slope(x, y) = 1.98$$

$$g(x) := intercept(x, y) + slope(x, y)\, x$$

Рис. 7.23. Фрагмент програми побудови лінійної однофакторної моделі за допомогою функцій intercept(*x*, *y*) та slope(*x*, *y*)

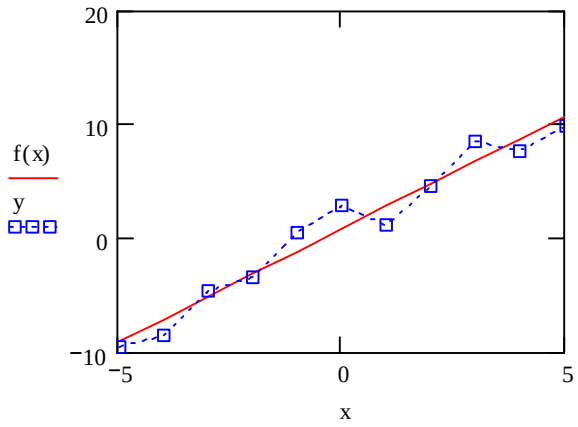


Рис. 7.24. Результат побудови лінійної однофакторної моделі

Функція line(x, y) видає вектор коефіцієнтів лінійної моделі $f(x) = ax+b$ у вигляді вектора $\begin{pmatrix} b \\ a \end{pmatrix}$.

Значенням функції intercept(x, y) є ордината точки перетину моделі з віссю ординат, а значенням функції slope(x, y) – тангенс кута нахилу моделі до осі абсцис.

На рис. 7.25 наведено фрагмент програми для побудови медіанної регресійної моделі за даними, що використовувалися при побудові лінійної моделі методом найменших квадратів.

$$\text{medfit}(x, y) = \begin{pmatrix} 0.867 \\ 2.1 \end{pmatrix}$$

$$q(x) := \text{medfit}(x, y)_0 + \text{medfit}(x, y)_1 \cdot x$$

Рис. 7.25. Фрагмент програми побудови медіанної регресійної моделі

Функція medfit(x, y) видає вектор коефіцієнтів лінійної моделі $f(x) = ax+b$ у вигляді вектора $\begin{pmatrix} b \\ a \end{pmatrix}$. На рис. 7.26 показано результати порівняння звичайної лінійної та медіанної моделей.

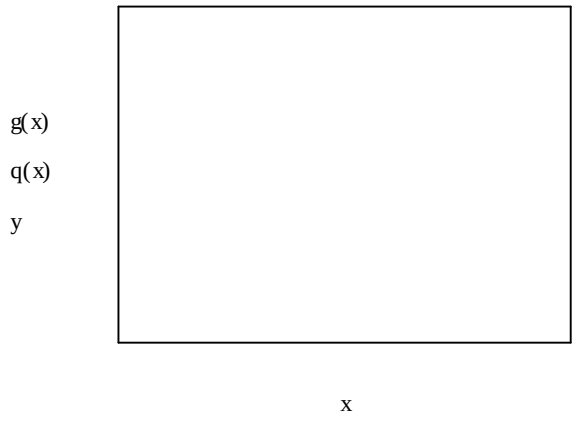


Рис. 7.26. Результат побудови лінійної однофакторної моделі

Для побудови поліноміальної моделі згенеруємо вектор $x1$, який містить 21 елемент, що утворюють арифметичну прогресію з початковим значенням –2,25 й різницею 0,25. Елементи вектора $y1$ розрахуємо за формулою:

$$y1 = -2x1^4 + 3x1^3 + 1,5x1^2 - 7x1 + 2 + \varepsilon,$$

де ε – елементи рівномірної випадкової послідовності, заданої на відрізку [–3; 3].

На рис. 7.27 наведено фрагмент програми, що використовується для побудови поліноміальної моделі.

```
x1 := (−2.5 −2.25 −2 −1.75 −1.5 −1.25 −1 −0.75 −0.5 −0.25 0 0.25 0.5 0.75 1 1.25 1.5 1.75 2 2.25 2.5)^T

y1 := (−96.8 −62.5 −33.8 −13.6 −2.3 5.1 2.6 5.6 7.5 1.6 2.8 −2.3 −3.7 −3.1 −4.2 −6.3 −6.4 −9.3 −13.7 −24.1 −38.1)^T

k := 4

s := regress(x1, y1, 4)

s = ( 3
      3
      4
      0.644
     −8.548
      2.366
      3.294
     −2.134 )

A(t) := interp(s, x1, y1, t)
```

Рис. 7.27. Фрагмент програми побудови поліноміальної регресійної моделі

Останні п’ять елементів вектора s є оцінками коефіцієнтів a_0, a_1, a_2, a_3 й a_4 полінома $A(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4$. На рис. 7.28 показано графік отриманої моделі, який достатньо добре узгоджується з вихідними даними.

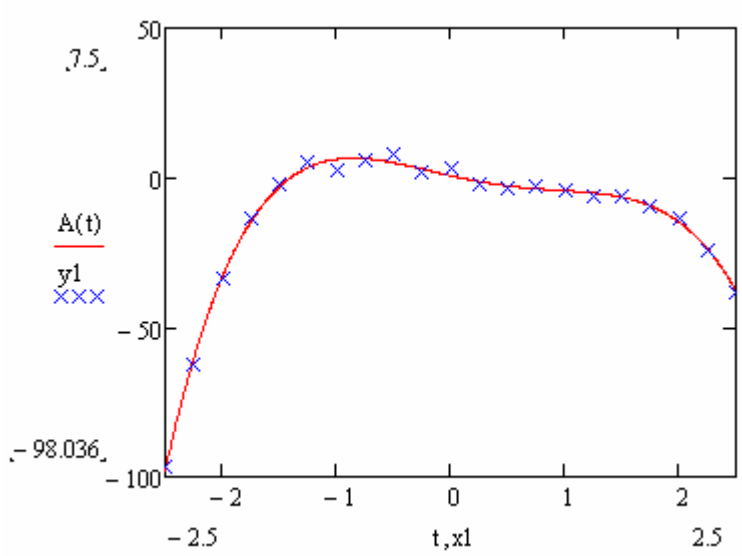


Рис. 7.28. Результат побудови регресійної моделі у вигляді полінома четвертого степеня

Для порівняння на рис. 7.29–7.32 показано графіки моделей у вигляді поліномів першого, другого, третього та п’ятого степенів, побудованих для тих самих вихідних даних.

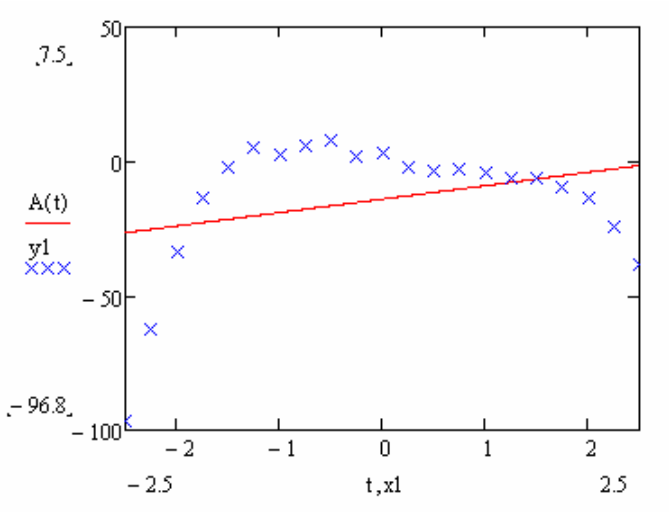


Рис. 7.29. Результат побудови регресійної моделі у вигляді полінома першого степеня (лінійної моделі)

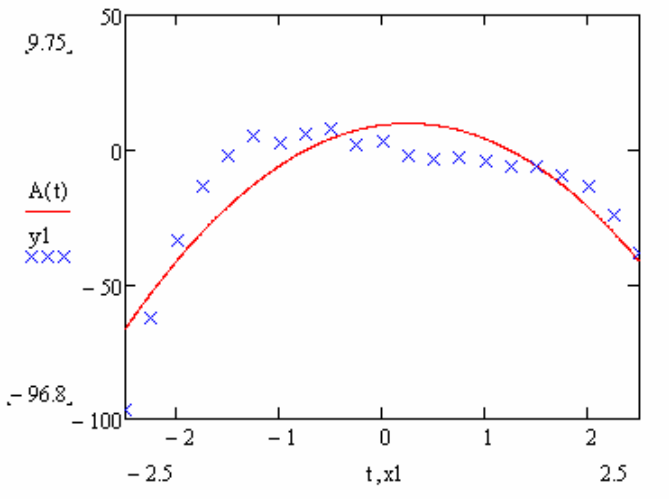


Рис. 7.30. Результат побудови регресійної моделі у вигляді полінома другого степеня (квадратичної моделі)

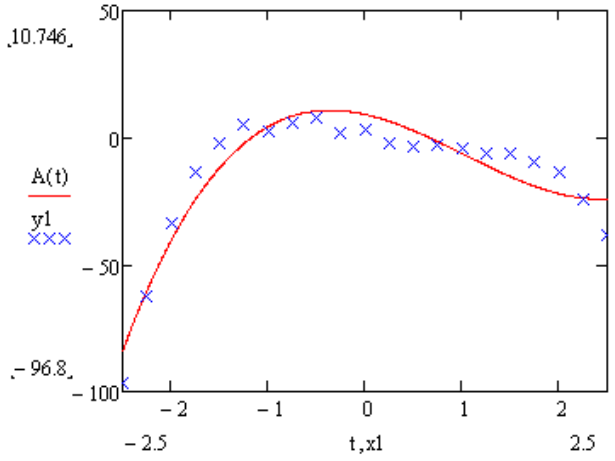


Рис. 7.31. Результат побудови регресійної моделі у вигляді полінома третього степеня (кубічної моделі)

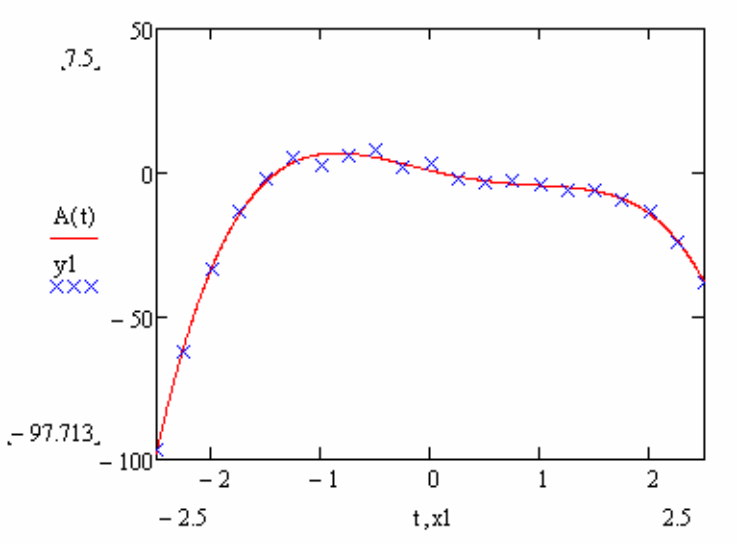


Рис. 7.32. Результат побудови регресійної моделі у вигляді полінома п’ятого степеня

Зіставлення наведених результатів дає підстави зробити висновок, що для наведених даних вже модель третього порядку достатньо добре відображає основні особливості вихідних даних, а модель п’ятого порядку є надлишковою, оскільки немає жодних переваг перед моделлю четвертого порядку, але є більш складною.

Альтернативний варіант передбачає побудову моделі у вигляді відрізків поліномів. Фрагмент програми для побудови такої моделі наведено на рис. 7.33. При цьому використовували ті самі дані, що і для побудови звичайної поліноміальної моделі.

```
s1 := loess(x1, y1, 0.75)

A1(t) := interp(s1, x1, y1, t)
```

Рис. 7.33. Фрагмент програми побудови регресійної моделі
у вигляді відрізків поліномів

На рис. 7.34–7.37 наведено результати побудови моделі для різних значень фактора span, що задає довжину відрізків поліномів. Аналіз наведених даних показує, що при малих значеннях фактора span модель краще відображає наявні дані. Але занадто добра відповідність моделі й вихідних даних може бути непотрібною, оскільки дані містять певну похибку.

Крім того для малих значень фактора span істотно збільшується обсяг розрахунків і при span $\leq$ 0,28 з’являється повідомлення про нестачу пам’яті для завершення операції. Із збільшенням параметра span понад 0,75–0,8 якість моделі погіршується, а при span = 1 відхилення моделі від даних вже є неприпустимо великими.

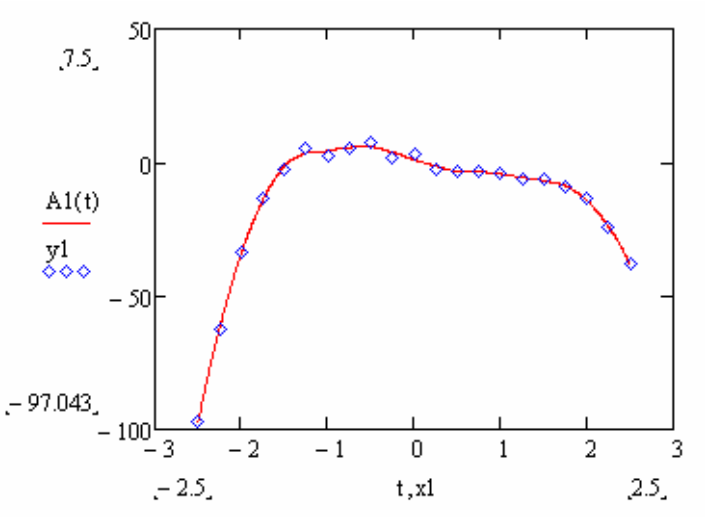


Рис. 7.34. Результат побудови регресійної моделі у відрізків поліномів для span = 0,3

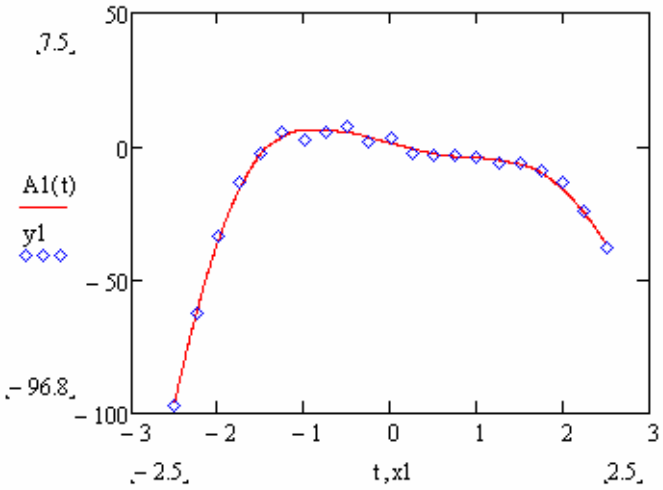


Рис. 7.35. Результат побудови регресійної моделі у відрізків поліномів для span = 0,5

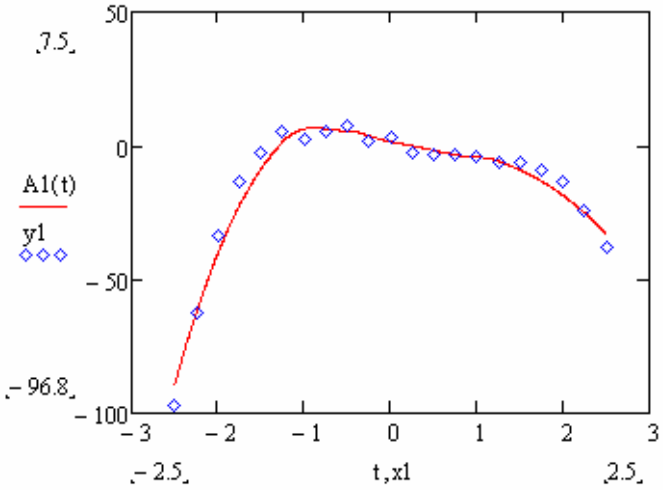


Рис. 7.36. Результат побудови регресійної моделі у відрізків поліномів для span = 0,75

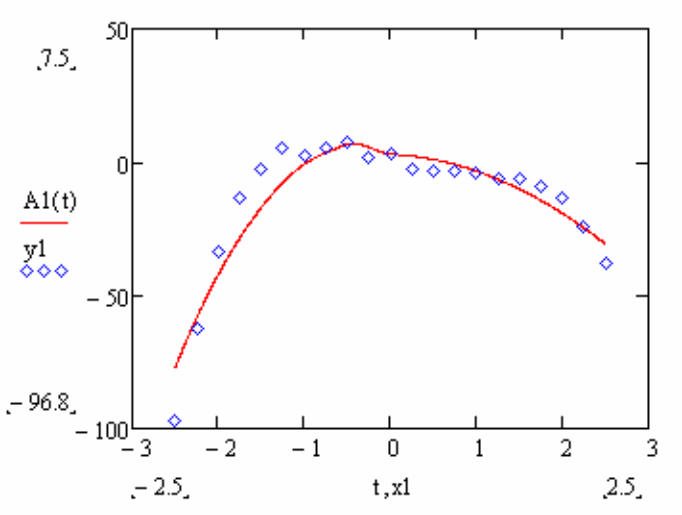


Рис. 7.37. Результат побудови регресійної моделі у відрізків поліномів для span = 1

Для побудови спеціальних моделей можна використовувати спеціальні функції пакету MathCad. Для ілюстрації цього візьмемо попередній вектор $x1$, а вектор $z1$ сформуємо за формулою:

$$z1 = \frac{5\varepsilon}{1 + 2\exp(-3x1)},$$

де ε – випадкова величина, рівномірно розподілена на відрізку [0,9; 1,1].

Для побудови експоненціальної та логістичної регресійних моделей використовували форму, фрагмент якої наведено на рис. 7.38.

$x1 := (-2.5 \ -2.25 \ -2 \ -1.75 \ -1.5 \ -1.25 \ -1 \ -0.75 \ -0.5 \ -0.25 \ 0 \ 0.25 \ 0.5 \ 0.75 \ 1 \ 1.25 \ 1.5 \ 1.75 \ 2 \ 2.25 \ 2.5)^T$

$z1 := (0.001 \ 0.003 \ 0.006 \ 0.014 \ 0.026 \ 0.061 \ 0.12 \ 0.27 \ 0.53 \ 0.88 \ 1.64 \ 2.34 \ 3.19 \ 3.92 \ 4.83 \ 5.16 \ 5.25 \ 5.37 \ 4.76 \ 5.09 \ 5.33)^T$

$$g := \begin{pmatrix} 5 \\ 2 \\ 2 \end{pmatrix}$$

$v := expfit(x1, z1, g)$

$$v = \begin{pmatrix} 10.233 \\ 0.135 \\ -8.123 \end{pmatrix}$$

$f(t) := v_0 \cdot \exp(v_1 \cdot t) + v_2$

$s := lgsfit(x1, z1, g)$

$$ff(t) := \frac{s_0}{1 + s_1 \cdot \exp(-s_2 \cdot t)} \qquad s = \begin{pmatrix} 5.266 \\ 2.461 \\ 2.865 \end{pmatrix}$$

Рис. 7.38. Фрагмент програми побудови експоненціальної та логістичної регресійної моделей

Результати побудови моделей показано на рис. 7.39, 7.40.

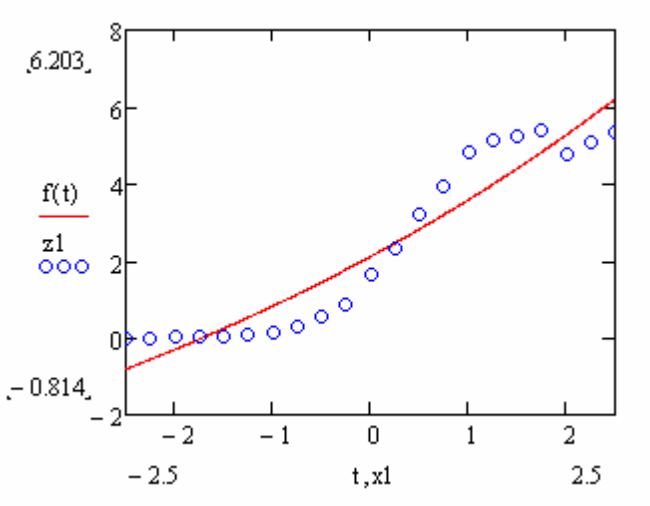


Рис. 7.39. Результат побудови експоненціальної регресійної моделі

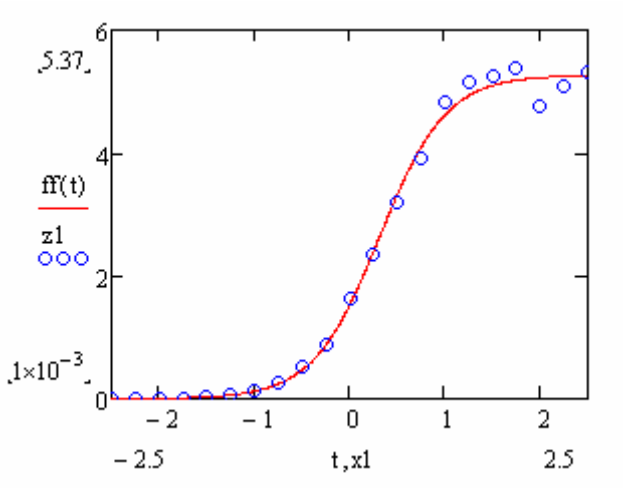


Рис. 7.40. Результат побудови логістичної регресійної моделі

Бачимо, що логістична модель значно краще відображає емпіричні дані, що у даному випадку є цілком природним оскільки вихідні дані побудовано саме на основі логістичної моделі. Дослідження впливу початкових значень параметрів моделей показує, що навіть для досить істотних їх відхилень від правильних значень, результат підбору параметрів моделей зазвичай є одним і тим самим. Але для окремих наборів вихідних значень алгоритм підбору параметрів не збігається.

На рис. 7.41 наведено фрагмент програми, яка дає змогу будувати регресійну модель у вигляді лінійної комбінації двох експонент.

Результат виконання цієї програми наведено на рис. 7.42. Бачимо, що побудована модель достатньо точно описує наведені дані.

Отримані результати дають підстави стверджувати, що пакет MathCad можна застосовувати для побудови однофакторних регресійних моделей різних типів.

$$x := (0\ \ 0.5\ \ 1\ \ 1.5\ \ 2\ \ 2.5\ \ 3\ \ 3.5\ \ 4\ \ 4.5\ \ 5\ \ 5.5\ \ 6\ \ 6.5\ \ 7\ \ 7.5\ \ 8\ \ 8.5\ \ 9\ \ 9.5\ \ 10)^T$$

$$y := (5.4\ \ 4.1\ \ 4.1\ \ 3.6\ \ 2.8\ \ 2.8\ \ 2.3\ \ 2.3\ \ 2\ \ 1.5\ \ 1.46\ \ 1.2\ \ 1.08\ \ 0.99\ \ 0.99\ \ 0.9\ \ 0.8\ \ 0.72\ \ 0.57\ \ 0.54\ \ 0.51)^T$$

$$F(x) := \begin{pmatrix} \exp\left(\frac{-x}{3}\right) \\ \exp\left(\frac{-x}{5}\right) \end{pmatrix}$$

$$C := \text{linfit}(x, y, F)$$

$$C = \begin{pmatrix} 2.343 \\ 2.79 \end{pmatrix}$$

$$f(x) := C_0 \exp\left(\frac{-x}{3}\right) + C_1 \exp\left(\frac{-x}{5}\right)$$

Рис. 7.41. Фрагмент програми побудови моделі у вигляді суми функцій

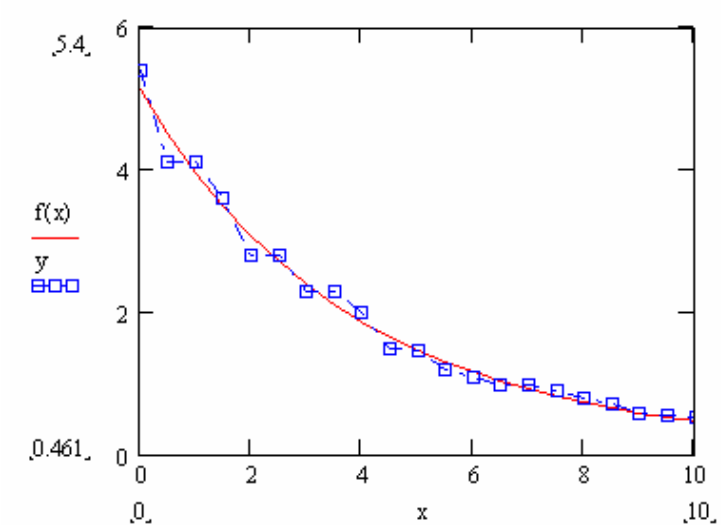


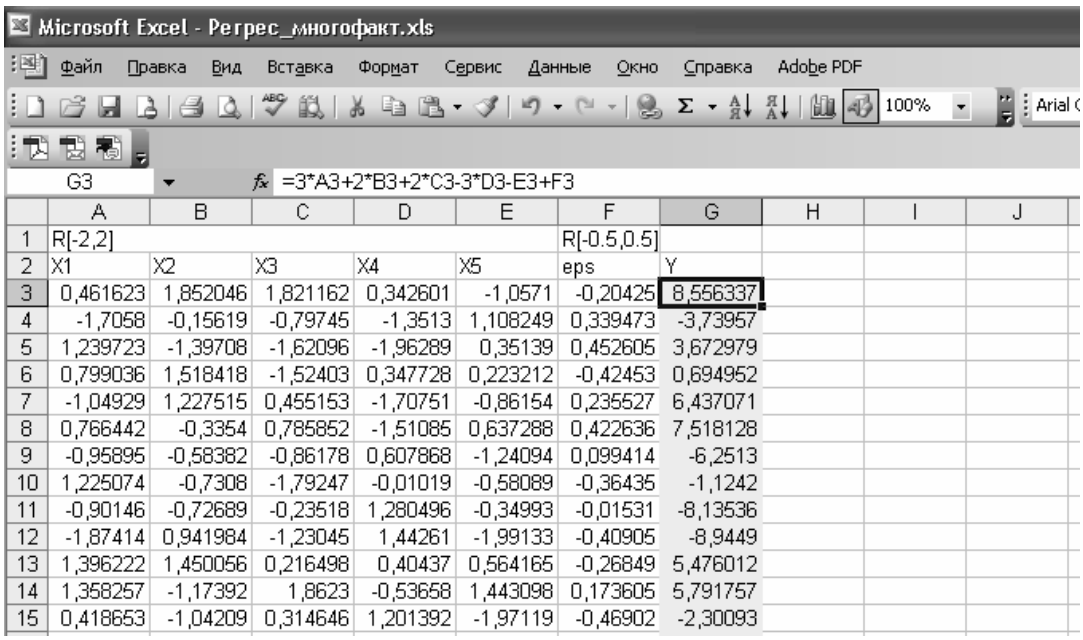
Рис. 7.42. Результат побудови моделі у вигляді суми двох експонент

7.11. Побудова лінійної багатофакторної моделі в електронних таблицях MS Excel

Створимо у на робочому аркуші масив, що містить по 50 значень п’яті змінних, що є елементами рівномірної випадкової послідовності, заданої на відрізку [2; 2]. Значення залежної змінної розрахуємо за формулою:

=3*A3+2*B3+2*C3-3*D3-E3+F3,

де A3–E3 – посилання на комірки зі значеннями сформованих змінних, а F3 – посилання на комірку, де міститься значення елементу рівномірної випадкової величини, яка задана на відрізку [–a; a]. Фрагмент робочого аркуша наведено на рис. 7.43.



	A	B	C	D	E	F	G	H	I	J
1	R[-2,2]					R[-0.5,0.5]				
2	X1	X2	X3	X4	X5	eps	Y			
3	0,461623	1,852046	1,821162	0,342601	-1,0571	-0,20425	8,556337			
4	-1,7058	-0,15619	-0,79745	-1,3513	1,108249	0,339473	-3,73957			
5	1,239723	-1,39708	-1,62096	-1,96289	0,35139	0,452605	3,672979			
6	0,799036	1,518418	-1,52403	0,347728	0,223212	-0,42453	0,694952			
7	-1,04929	1,227515	0,455153	-1,70751	-0,86154	0,235527	6,437071			
8	0,766442	-0,3354	0,785852	-1,51085	0,637288	0,422636	7,518128			
9	-0,95895	-0,58382	-0,86178	0,607868	-1,24094	0,099414	-6,2513			
10	1,225074	-0,7308	-1,79247	-0,01019	-0,58089	-0,36435	-1,1242			
11	-0,90146	-0,72689	-0,23518	1,280496	-0,34993	-0,01531	-8,13536			
12	-1,87414	0,941984	-1,23045	1,44261	-1,99133	-0,40905	-8,9449			
13	1,396222	1,450056	0,216498	0,40437	0,564165	-0,26849	5,476012			
14	1,358257	-1,17392	1,8623	-0,53658	1,443098	0,173605	5,791757			
15	0,418653	-1,04209	0,314646	1,201392	-1,97119	-0,46902	-2,30093			

Рис. 7.43. Фрагмент робочого аркуша з даними

Для побудови лінійної регресійної моделі скористуємося засобом “Регресія” Пакету аналізу електронних таблиць MS Excel. У діалоговому вікні (рис. 7.44) помічаємо комірки, що містять незалежні й залежну змінні, а також які саме результати потрібно вивести.

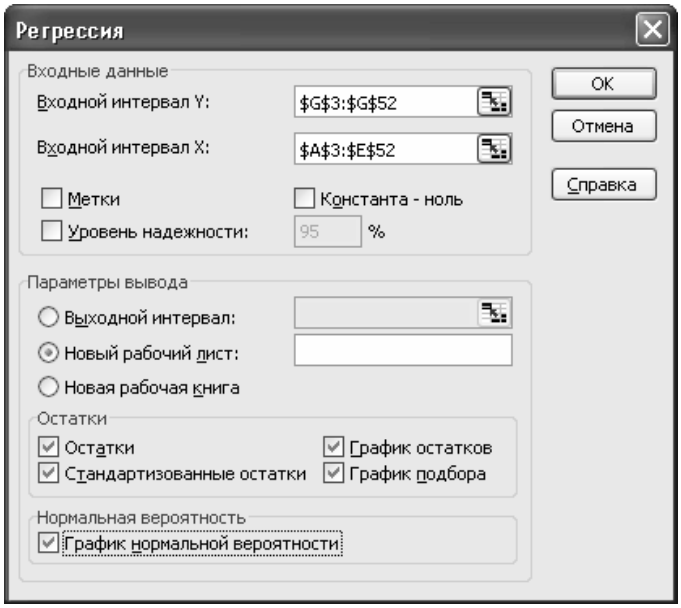


Рис. 7.44. Діалогове вікно підбору параметрів регресійної моделі

Результати роботи програми наведено на рис. 7.45. Звідси бачимо, що модель можна записати у вигляді:

$$Y = 2,98x_1 + 2,02x_2 + 1,99x_3 - 3,05x_4 - 0,96x_5 .$$

Регрессионная статистика								
Множественный R	0,998815							
R-квадрат	0,997631							
Нормированный R-квадрат	0,997362							
Стандартная ошибка	0,327466							
Наблюдения	50							
Дисперсионный анализ								
	df	*SS*	*MS*	*F*	*ачимость F*			
Регрессия	5	1987,279	397,4558	3706,431	1,46E-56			
Остаток	44	4,7183	0,107234					
Итого	49	1991,997						
	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	0,003106	0,047113	0,065927	0,947734	-0,09184	0,098055	-0,09184	0,098055
Переменная X 1	2,98268	0,043683	68,2808	2,73E-46	2,894643	3,070716	2,894643	3,070716
Переменная X 2	2,020953	0,04242	47,64104	1,67E-39	1,935461	2,106446	1,935461	2,106446
Переменная X 3	1,98938	0,044526	44,67861	2,66E-38	1,899642	2,079117	1,899642	2,079117
Переменная X 4	-3,04665	0,046941	-64,9033	2,49E-45	-3,14125	-2,95204	-3,14125	-2,95204
Переменная X 5	-0,9552	0,041872	-22,8127	5,19E-26	-1,03959	-0,87082	-1,03959	-0,87082

Рис. 7.45. Результати підбору лінійної моделі

Стандартне відхилення вільного члена істотно перевищує його значення, тому включати вільний член до моделі недоцільно. Коефіцієнт детермінації побудованої моделі дорівнює 0,997 і є досить близьким до одиниці, що свідчить про адекватність лінійної моделі.

Графіки залишків (рис. 7.46) вказують на їх випадковий характер, що також є свідченням адекватності моделі.

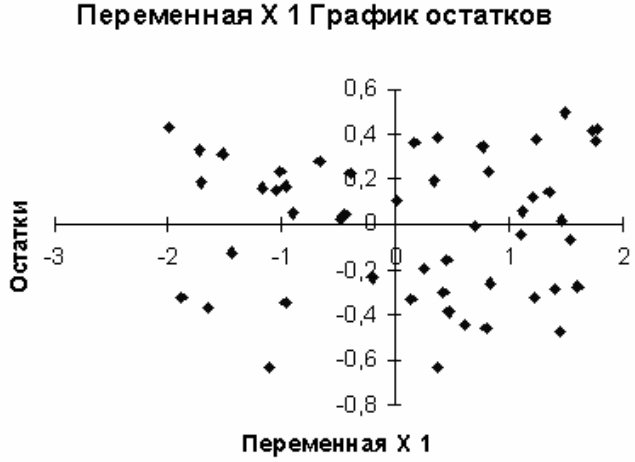


Рис. 7.46. Графік залишків для змінної x_1

Процедура “Регресія” електронних таблиць MS Excel також дає змогу побудувати графіки підбору, що характеризують відповідність моделі вихідним даним. Для досліджуваної моделі вони мають вигляд, наведений на рис. 7.47.

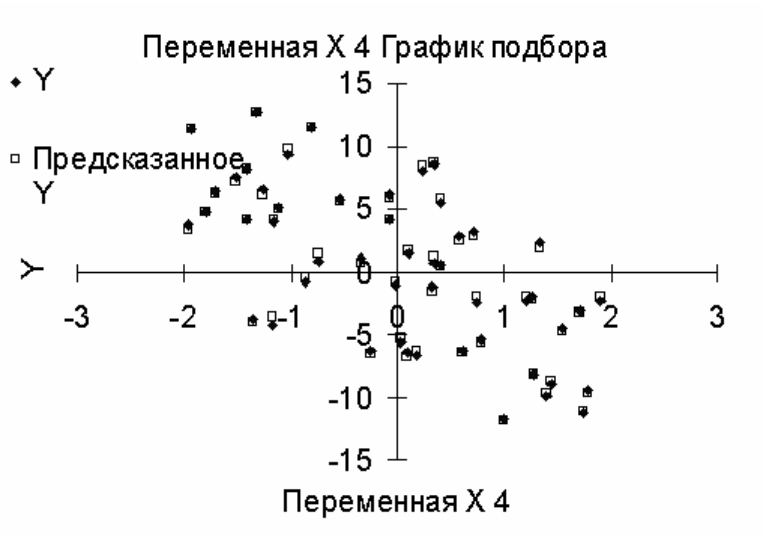


Рис. 7.47. Графік підбору для змінної x_4

На рис. 7.48 показано графік нормального розподілу, який дає змогу перевірити відповідність ряду залишків нормальному розподілу. Але цей графік не зручний для практичного застосування.

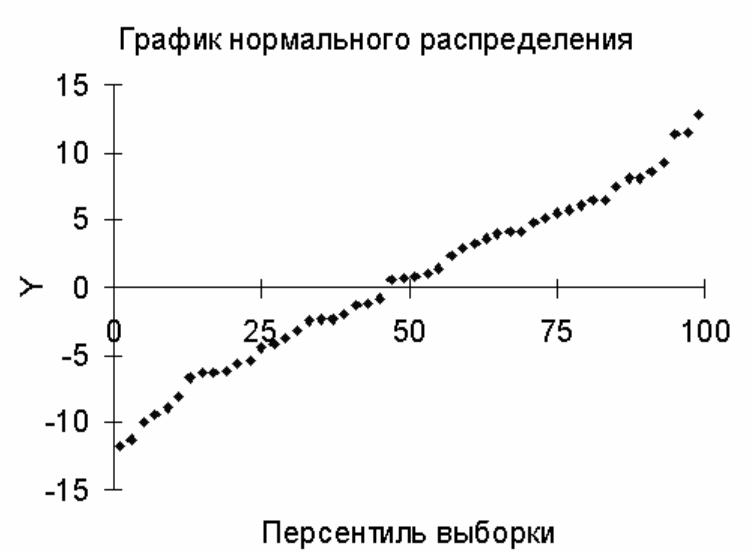
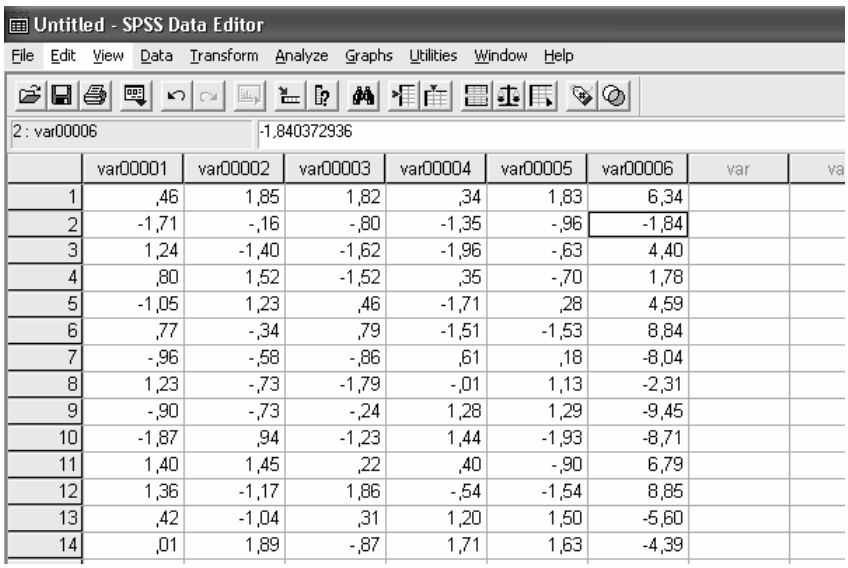


Рис. 7.48. Графік нормального розподілу

7.12. Побудова лінійної багатофакторної моделі в пакеті SPSS

Для побудови й дослідження багатофакторних регресійних моделей також можна також використовувати статистичний пакет SPSS [17]. Проілюструємо це на прикладі розглянутої вище моделі з п’ятма незалежними змінними.

Заносимо на робочий аркуш значення усіх змінних (рис. 7.49)



Untitled - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

2 : var00006 -1,840372936

	var00001	var00002	var00003	var00004	var00005	var00006	var	va
1	,46	1,85	1,82	,34	1,83	6,34		
2	-1,71	-,16	-,80	-1,35	-,96	-1,84		
3	1,24	-1,40	-1,62	-1,96	-,63	4,40		
4	,80	1,52	-1,52	,35	-,70	1,78		
5	-1,05	1,23	,46	-1,71	,28	4,59		
6	,77	-,34	,79	-1,51	-1,53	8,84		
7	-,96	-,58	-,86	,61	,18	-8,04		
8	1,23	-,73	-1,79	-,01	1,13	-2,31		
9	-,90	-,73	-,24	1,28	1,29	-9,45		
10	-1,87	,94	-1,23	1,44	-1,93	-8,71		
11	1,40	1,45	,22	,40	-,90	6,79		
12	1,36	-1,17	1,86	-,54	-1,54	8,85		
13	,42	-1,04	,31	1,20	1,50	-5,60		
14	,01	1,89	-,87	1,71	1,63	-4,39		

Рис. 7.49. Вигляд робочого аркушу SPSS з вихідними даними

Потім у головному меню обираємо Analyze/Regression/Linear regression. Після цього з'являється вікно вибору параметрів моделі (рис. 7.50). У цьому вікні позначаємо залежну й незалежні змінні. У вікні Statistics (рис. 7.51) позначаємо, які параметри моделі необхідно вивести. Зокрема можна передбачити вивід довірчих інтервалів для коефіцієнтів моделі, перевірку мультиколінеарності даних, перевірку автокореляції залишків моделі за критерієм Дарбіна-Уотсона тощо.

У вікні Plots (рис. 7.52) позначаємо які графіки необхідно побудувати й які змінні відкладати за їх осями. Зокрема ми можемо побудувати гістограму залишків і відповідний графік нормального розподілу. У вікні Save (рис. 7.53) зазначаємо, які змінні необхідно додати у вікні даних.

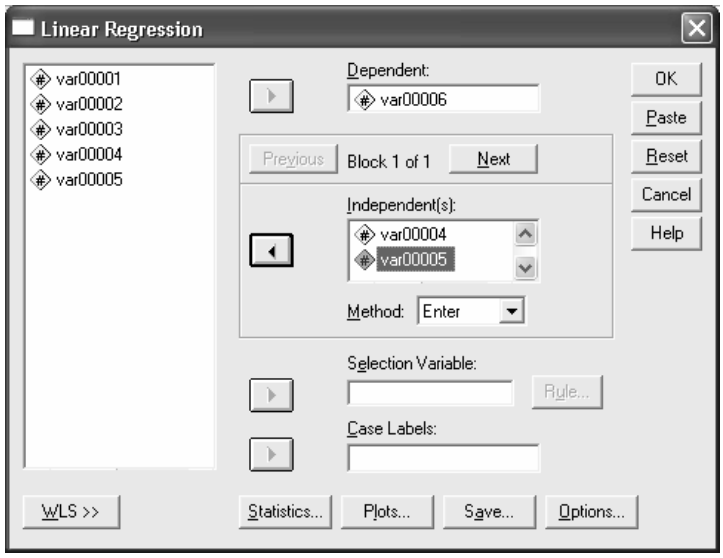


Рис. 7.50. Діалогове вікно вибору параметрів побудови регресійної моделі

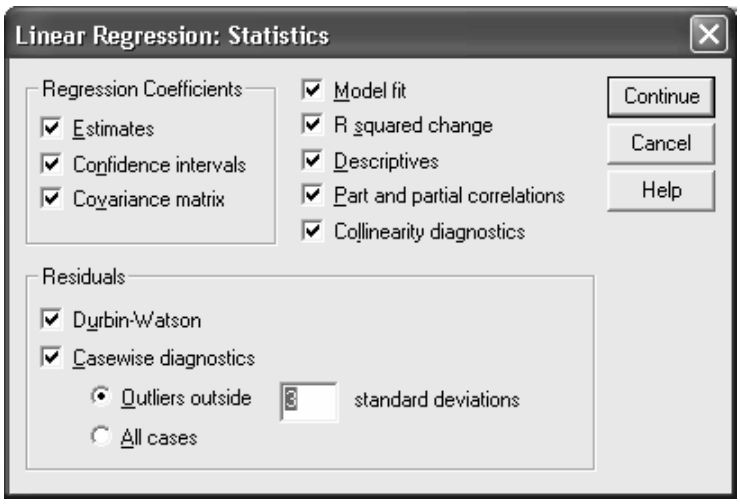


Рис. 7.51. Діалогове вікно задання статистичних параметрів моделі, які необхідно вказати у вікні виводу

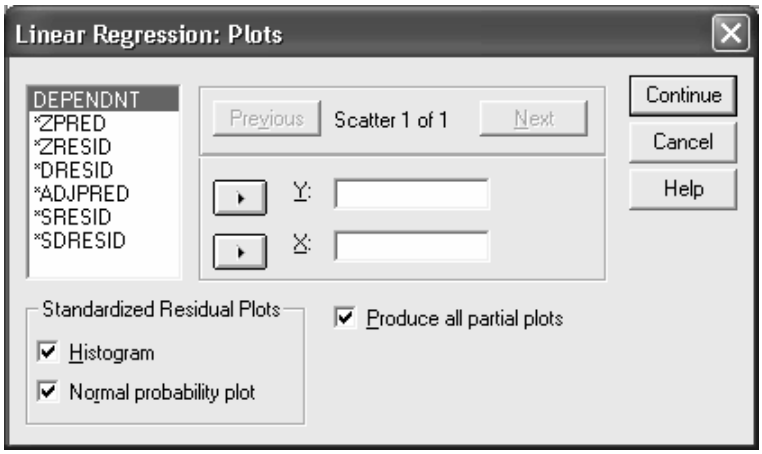


Рис. 7.52. Діалогове вікно задання параметрів графіків

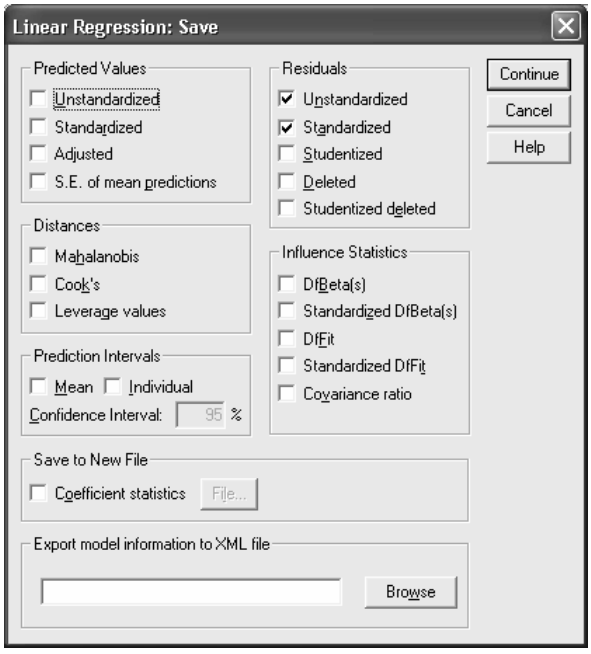


Рис. 7.53. Діалогове вікно задання величин, значення яких треба додати у вікні даних

У вікні Options (рис. 7.54) зазначаємо метод і параметри розрахункової процедури.

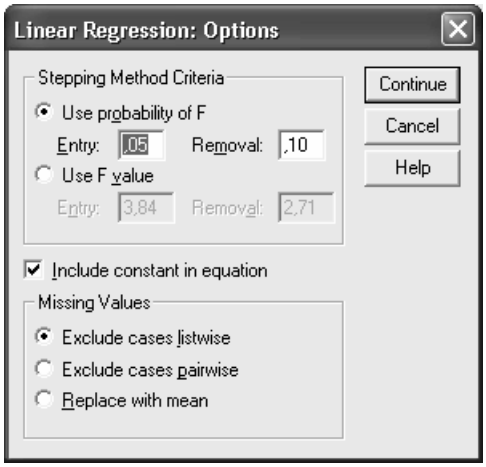


Рис. 7.54. Вікно задання методу і параметрів розрахункової процедури

Деякі результати підбору параметрів та аналізу багатофакторної лінійної моделі наведено на рис. 7.55–7.63.

Correlations

		VAR00006	VAR00001	VAR00002	VAR00003	VAR00004	VAR00005
Pearson Correlation	VAR00006	1,000	,712	,344	,480	-,664	-,312
	VAR00001	,712	1,000	,171	,097	-,192	-,237
	VAR00002	,344	,171	1,000	-,131	,045	,041
	VAR00003	,480	,097	-,131	1,000	-,276	-,057
	VAR00004	-,664	-,192	,045	-,276	1,000	,040
	VAR00005	-,312	-,237	,041	-,057	,040	1,000
Sig. (1-tailed)	VAR00006	.	,000	,007	,000	,000	,014
	VAR00001	,000	.	,118	,251	,090	,049
	VAR00002	,007	,118	.	,182	,377	,388
	VAR00003	,000	,251	,182	.	,026	,347
	VAR00004	,000	,090	,377	,026	.	,391
	VAR00005	,014	,049	,388	,347	,391	.
N	VAR00006	50	50	50	50	50	50
	VAR00001	50	50	50	50	50	50
	VAR00002	50	50	50	50	50	50
	VAR00003	50	50	50	50	50	50
	VAR00004	50	50	50	50	50	50
	VAR00005	50	50	50	50	50	50

Рис. 7.55. Кореляції між змінними

Зокрема з рис. 7.55 бачимо, що немає істотної кореляції між незалежними змінними VAR0001 – VAR0005.

Descriptive Statistics

	Mean	Std. Deviation	N
VAR00006	,3436	6,92320	50
VAR00001	,1572	1,13998	50
VAR00002	,0651	1,14300	50
VAR00003	-,1018	1,10601	50
VAR00004	-,0062	1,12834	50
VAR00005	,0939	1,19823	50

Рис. 7.56. Описова статистика для змінних

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	VAR00005, VAR00004, VAR00002, VAR00003, VAR00001[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: VAR00006

Рис. 7.57. Результати відбору незалежних змінних,
які враховуватимуться при побудові моделі

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,999[a]	,998	,998	,28419	,998	5807,296	5	44	,000	1,537

a. Predictors: (Constant), VAR00005, VAR00004, VAR00002, VAR00003, VAR00001

b. Dependent Variable: VAR00006

Рис. 7.58. Загальна характеристика моделі

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2345,048	5	469,010	5807,296	,000[a]
	Residual	3,554	44	,081		
	Total	2348,601	49			

a. Predictors: (Constant), VAR00005, VAR00004, VAR00002, VAR00003, VAR00001

b. Dependent Variable: VAR00006

Рис. 7.59. Таблиця ANOVA побудованої моделі

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	3,091E-02	,041		,753	,456	-,052	,114
	VAR00001	2,970	,038	,489	77,836	,000	2,893	3,047
	VAR00002	2,020	,037	,333	55,150	,000	1,946	2,094
	VAR00003	2,077	,039	,332	53,798	,000	1,999	2,154
	VAR00004	-2,984	,038	-,486	-78,425	,000	-3,061	-2,908
	VAR00005	-,988	,035	-,171	-28,225	,000	-1,059	-,918

a. Dependent Variable: VAR00006

Рис. 7.60. Таблиця коефіцієнтів побудованої моделі

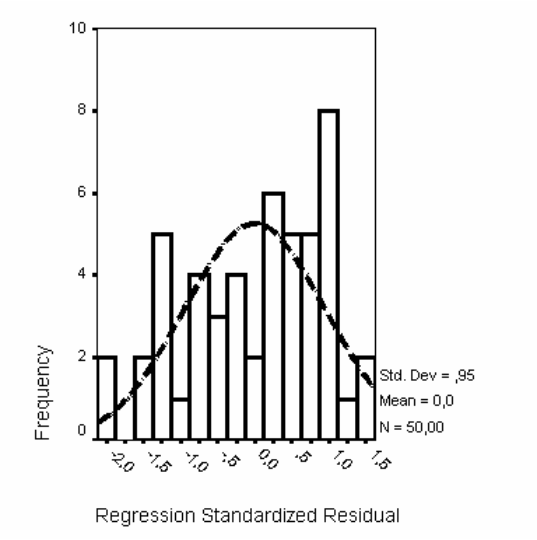


Рис. 7.61. Гістограма залишків

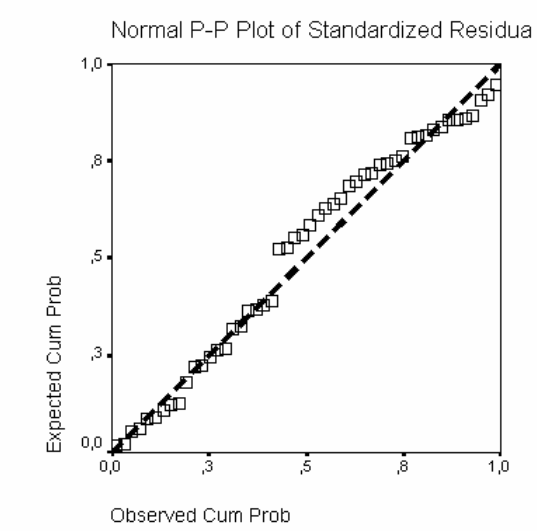


Рис. 7.62. P-P графік стандартизованих залишків моделі для нормального розподілу

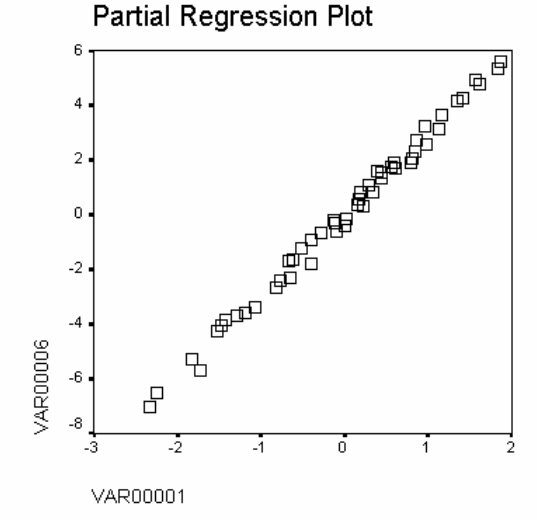


Рис. 7.63. Частинна регресія стосовно першої незалежної змінної

Побудована модель є адекватною, про що свідчить близькість отриманих значень коефіцієнтів, високий коефіцієнт детермінації й практична відсутність автокореляції залишків. Але гістограма залишків дещо відрізняється від нормального закону. Графіки частинної кореляції свідчать про гарне наближення до лінійного зв'язку за усіма незалежними змінними.

З наведених даних можна зробити висновок, що статистичний пакет SPSS надає більше можливостей для аналізу багатофакторних лінійних регресійних моделей, ніж електронні таблиці MS Excel, а одержувані за його допомогою результати є більш зручними для практичного використання.

Контрольні питання

1. Яким є основне завдання регресійного аналізу?
2. У чому полягають основні припущення класичного регресійного аналізу?
3. Якою є звичайна процедура класичного регресійного аналізу?
4. Як формулюється задача побудови регресійної моделі?
5. Які функціонали використовують для визначення параметрів регресійних моделей? У чому полягають переваги й недоліки різних типів таких функціоналів?
6. Якими є основні типи функцій, що використовуються для побудови однофакторних регресійних моделей?
7. Які моделі називають лінійними? Що називають порядком регресійної моделі?
8. Чому регресійні моделі не рекомендують використовувати поза межами тієї області значень вихідних параметрів, для якої вони побудовані?
9. Для заданого набору даних побудувати однофакторну лінійну регресійну модель і перевірити її адекватність.
10. У яких випадках нелінійні однофакторні моделі можна звести до лінійних? Навести приклади відповідних перетворень.
11. Для заданого набору даних побудуйте однофакторну нелінійну регресійну модель і перевірте її адекватність.
12. Як використовують критерій Фішера для перевірки адекватності регресійних моделей?
13. Як визначають довірчі інтервали для коефіцієнтів однофакторних регресійних моделей?
14. Яким є загальний вигляд поліноміальної регресійної моделі?
15. Яким є загальний алгоритм визначення порядку і параметрів поліноміальних регресійних моделей?
16. Для заданого набору даних побудуйте поліноміальну регресійну модель і перевірте її адекватність.

17. У яких випадках використовують регресійні моделі у вигляді тригонометричних поліномів? Яким є загальний алгоритм побудови таких моделей?

18. Для заданого набору даних побудуйте регресійну модель у вигляді тригонометричного поліному і перевірте її адекватність.

19. Якими є загальні алгоритми побудови однофакторних регресійних моделей у вигляді модифікованої показникової функції, кривої Гомперця та логістичної кривої?

20. Яким є загальний алгоритм побудови багатофакторної лінійної регресійної моделі?

21. Для заданого набору даних побудуйте багатофакторну лінійну регресійну модель і перевірте її адекватність.

22. Що називають мультиколінеарністю даних? Наведіть приклади.

23. Для чого застосовують алгоритми зміщеного оцінювання параметрів багатофакторних лінійних регресійних моделей? Наведіть приклади.

24. За якими властивостями перевіряють адекватність регресійних моделей? Якими є основні критерії адекватності?