Лекція 02.
Лінійні регресійні моделі

З метою досліджень часто буває зручно представити досліджуваний об'єкт у вигляді ящика, що має входи й виходи, не розглядаючи детально його внутрішньої структури. Звичайно, перетворення в ящику (на об'єкті) відбуваються (сигнали проходять по зв'язках й елементам, міняють свою форму й т.п.), але при такому поданні вони відбуваються приховано від спостерігача.

По ступені інформованості дослідника про об'єкт існує ділення об'єктів на три типи «ящиків»:

- «білий ящик»: про об'єкт відомо все;
- «сірий ящик»: відомий структура об'єкта, невідомі кількісні значення параметрів;
- «чорний ящик»: про об'єкт невідомо нічого.

Чорний ящик умовно зображують як на мал. 2.1.

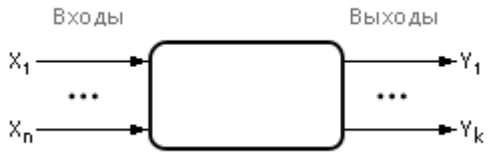

Рис. 2.1. Позначення чорного ящика на схемах

Значення на входах і виходах чорного ящика можна спостерігати й вимірювати. Уміст ящика невідомо.

Завдання полягає в тім, щоб, знаючи множину значень на входах і виходах, побудувати модель, тобто визначити функцію ящика, по якій вхід перетвориться у вихід. Така задача називається задачею регресійного аналізу.

Залежно від того, доступні входи дослідникові для керування або тільки для спостереження, можна говорити про активний або пасивний експеримент із ящиком.

Нехай, наприклад, перед нами постає задача визначити, як залежить випуск продукції від кількості споживаної електроенергії. Результати спостережень відобразимо на графіку (див. мал. 2.2). Усього на графіку n експериментальних крапок, які відповідають n спостереженням.

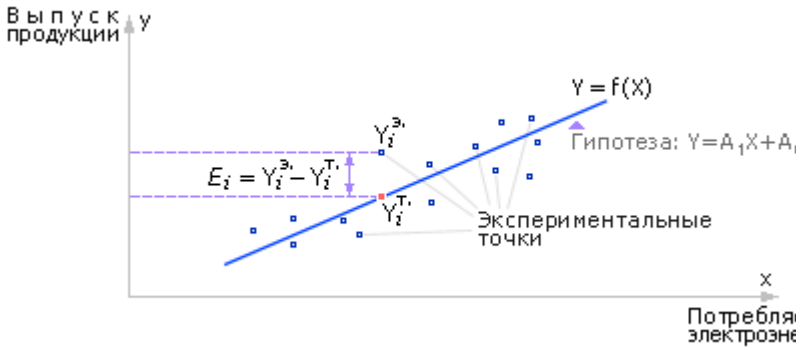

Рис. 2.2. Графічний вид подання результатів
спостереження над чорним ящиком

Для початку припустимо, що ми маємо справу із чорним ящиком, що має один вхід й один вихід. Допустимо для простоти, що залежність між входом і виходом лінійна або майже лінійна. Тоді дана модель буде називатися лінійною одномірною регресійною моделлю.

1) Дослідник вносить гіпотезу про структуру ящика

Розглядаючи експериментально отримані дані, припустимо, що вони підкоряються лінійній гіпотезі, тобто вихід Y залежить від входу X лінійно, тобто гіпотеза має вигляд: $Y = A_1X + A_0$ (мал. 2.2).

2) Визначення невідомих коефіцієнтів A_0 й A_1 моделі

Лінійна одномірна модель (мал. 2.3).

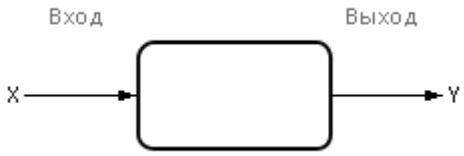

Рис. 2.3. Одномірна модель чорного ящика

Для кожної з n знятих експериментально крапок обчислимо помилку (E_i) між експериментальним значенням ($Y_i^{\text{Эксп.}}$) і теоретичним значенням ($Y_i^{\text{Теор.}}$), що лежить на гіпотетичній прямій $A_1X + A_0$ (див. мал. 2.2):

$$E_i = (Y_i^{\text{Эксп.}} - Y_i^{\text{Теор.}}),\ i = 1, \ldots, n;$$

$$E_i = Y_i - A_0 - A_1 \cdot X_i,\ i = 1, \ldots, n.....$$

Помилки E_i для всіх n крапок варто скласти. Щоб позитивні помилки не компенсували в сумі негативні, кожну з помилок підносять до квадрата й складають їхні значення в сумарну помилку F уже одного знака:

$$E_i^2 = (Y_i - A_0 - A_1 \cdot X_i)^2,\ i = 1, \ldots, n...$$

$$F(A_0,\ A_1) = \sum_{i=1}^{n} E_i^2$$

Ціль методу — мінімізація сумарної помилки F за рахунок підбора коефіцієнтів A_0, A_1. Інакше кажучи, це означає, що необхідно знайти такі коефіцієнти A_0, A_1 лінійної функції $Y = A_1 X + A_0$, щоб її графік проходив якнайближче одночасно до всіх експериментальних крапок. Тому даний метод називається методом найменших квадратів.

$$F(A_0,\ A_1) = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (Y_i - A_0 - A_1 X_i)^2 \Rightarrow \min_{A_0,\ A_1}$$

Сумарна помилка F є функцією двох змінних A_0 й A_1, тобто $F(A_0, A_1)$, міняючи які, можна впливати на величину сумарної помилки (див. мал. 2.4).

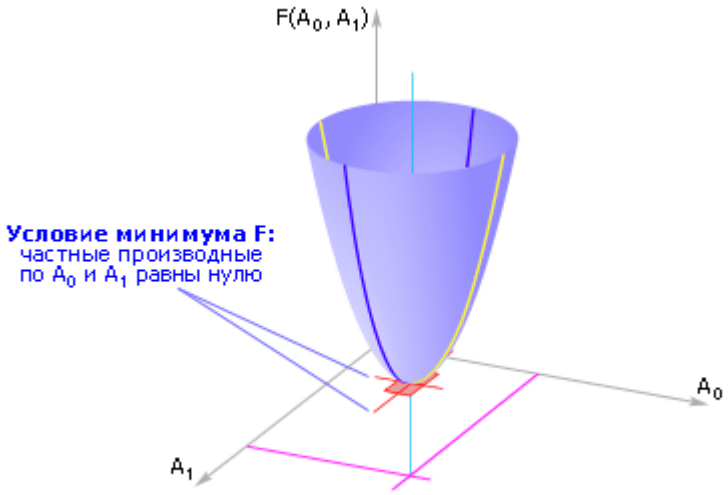

Рис. 2.4. Зразковий вид функції помилки

Щоб сумарну помилку мінімізувати, знайдемо частки похідні від функції F по кожної змінній і дорівняємо їх до нуля (умова екстремума):

$$\frac{\partial F}{\partial A_0} = -2 \sum_{i=1}^{n} (Y_i - A_0 - A_1 X_i) = 0$$

$$\frac{\partial F}{\partial A_1} = -2 \sum_{i=1}^{n} (Y_i - A_0 - A_1 X_i) X_i = 0$$

Після розкриття дужок одержимо систему із двох лінійних рівнянь:

$$\sum_{i=1}^{n} A_0 + A_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i$$

$$A_0 \sum_{i=1}^{n} X_i + A_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i$$

Для знаходження коефіцієнтів A_0 й A_1 методом Крамера представимо систему в матричній формі:

$$\begin{pmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \end{pmatrix} \cdot \begin{pmatrix} A_0 \\ A_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_i Y_i \end{pmatrix}$$

Рішення має вигляд:

$$A_0 = \frac{\sum_{i=1}^{n} Y_i \sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} X_i Y_i \sum_{i=1}^{n} X_i}{n \sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}$$

$$A_1 = \frac{n \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} Y_i \sum_{i=1}^{n} X_i}{n \sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}$$

Обчислюємо значення A_0 й A_1.

3) Перевірка

Щоб визначити, приймається чи гіпотеза ні, потрібно, по-перше, розрахувати помилку між крапками заданої експериментальної й отриманої теоретичної залежності й сумарну помилку:

$$E_i = (Y_i^{\text{Эксп.}} - Y_i^{\text{Теор.}}),\ i = 1, \ldots, n$$

$$F(A_0, A_1) = \sum_{i=1}^{n} E_i^2$$

І, по-друге, необхідно знайти значення σ по формулі $\sigma = \sqrt{\frac{F}{n}}$, де F — сумарна помилка, n — загальне число експериментальних крапок.

Якщо в смугу, обмежену лініями $Y^{\text{Теор.}} - S$ й $Y^{\text{Теор.}} + S$ (мал. 2.5), попадає 68.26% і більше експериментальних крапок $Y_i^{\text{Эксп.}}$, те висунута нами гіпотеза приймається. У противному випадку вибирають більше складну гіпотезу або перевіряють вихідні дані. Якщо потрібна більша впевненість у результаті, то використають додаткову умову: у смугу, обмежену лініями $Y^{\text{Теор.}} - 2S$ й $Y^{\text{Теор.}} + 2S$, повинні потрапити 95.44% і більше експериментальних крапок $Y_i^{\text{Эксп.}}$.

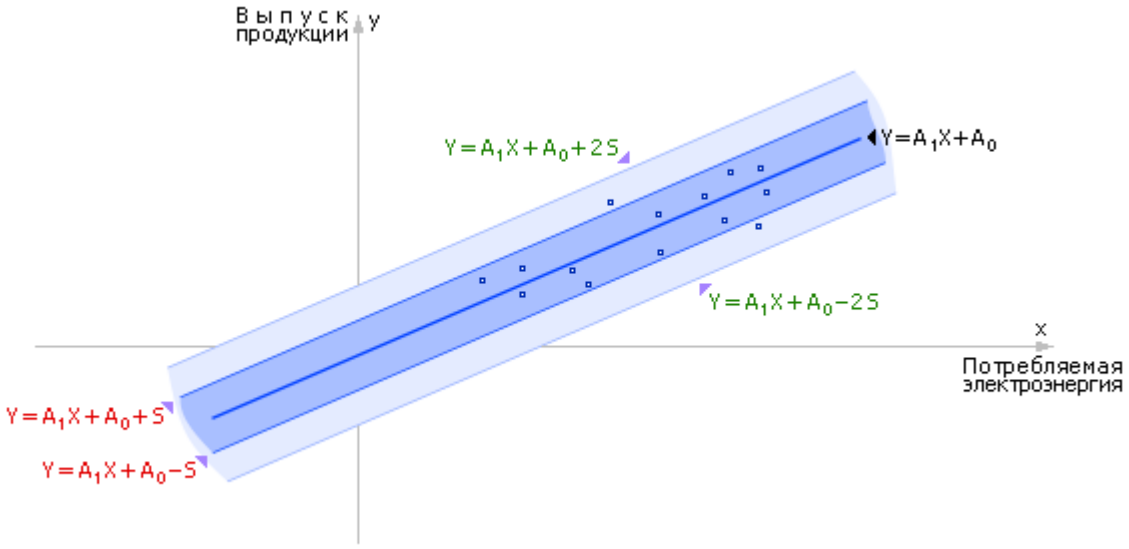

Рис. 2.5. Дослідження допустимості прийняття гіпотези

Відстань S пов'язане з σ наступним співвідношенням:

$$S = \sigma/\sin(\beta) = \sigma/\sin(90° - \text{arctg}(A_1)) = \sigma/\cos(\text{arctg}(A_1)),$$

що проілюстровано на мал. 2.6.

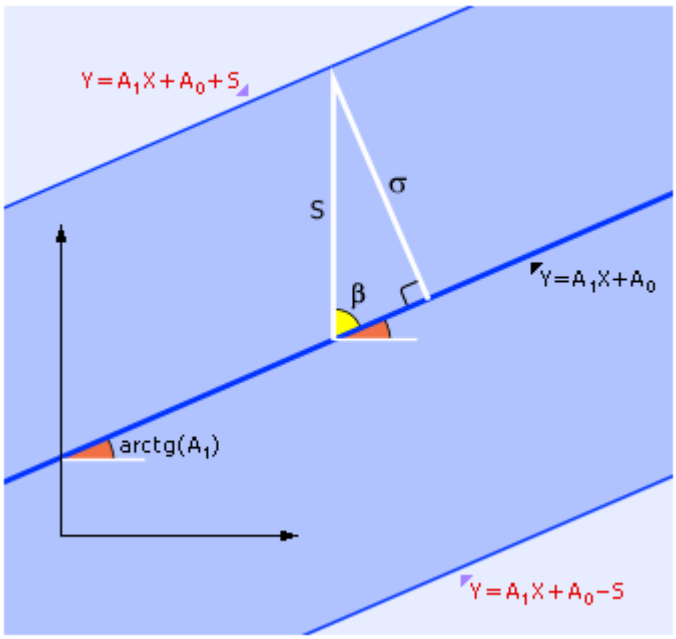

Рис. 2.6. Зв'язок значень σ і S

Умова прийняття гіпотези виведено з нормального закону розподілу випадкових помилок (див. мал. 2.7). P — імовірність розподілу нормальної помилки.

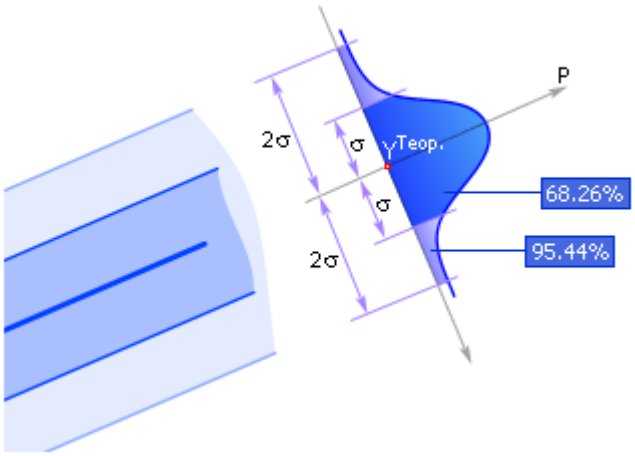

Рис. 2.7. Ілюстрація закону
нормального розподілу помилок

Нарешті, приведемо на мал. 2.8 графічну схему реалізації одномірної лінійної регресійної моделі.

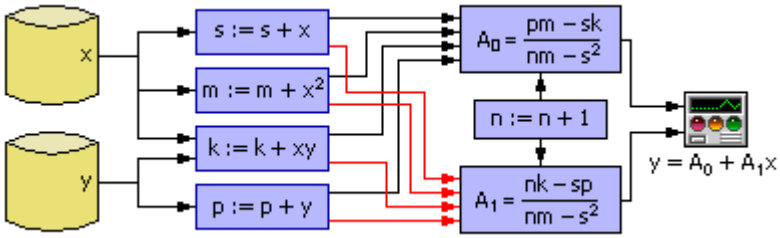

Рис. 2.8. Схема реалізації методу
найменших квадратів у середовищі моделювання

Лабораторна модель

Припустимо, що функціональна структура ящика знову має лінійну залежність, але кількість вхідних сигналів, що діють одночасно на об'єкт, дорівнює *m* (див. мал. 2.9):

$$Y = A_0 + A_1 \cdot X_1 + \ldots + A_m \cdot X_m \ldots$$

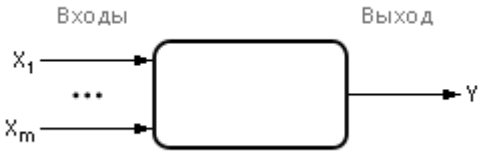

Рис. 2.9. Позначення багатомірного
чорного ящика на схемах

Тому що мається на увазі, що ми маємо експериментальні дані про всі входи й виходи чорного ящика, то можна обчислити помилку між експериментальним ($Y_i^{\text{Эксп.}}$) і теоретичним ($Y_i^{\text{Teop.}}$) значенням *Y* для кожної *i-oї* крапки (нехай, як і колись, число експериментальних крапок дорівнює *n*):

$$E_i = (Y_i^{\text{Эксп.}} - Y_i^{\text{Теор.}}),\ i = 1, \ldots, n;$$

$$E_i = Y_i - A_0 - A_1 \cdot X_{1i} - \ldots - A_m \cdot X_{mi},\ i = 1, \ldots, n.....$$

Мінімізуємо сумарну помилку F:

$$F(A_0, A_1, \ldots, A_m) = \sum_{i=1}^{n} E_i^2 \Rightarrow \min_{A_0, A_1, \ldots, A_m}$$

Помилка F залежить від вибору параметрів A_0, A_1, …, A_m... Для знаходження экстремума дорівняємо всі частки похідні F по невідомим $A_0, A_1, \ldots, A_m$ до нуля:

$$\frac{\partial F}{\partial A_j} = 0,\ j = \overline{0,\ m}$$

Одержимо систему з $m+1$ рівняння з $m+1$ невідомими, котру варто вирішити, щоб визначити коефіцієнти лінійної множинної моделі A_0, A_1, …, A_m... Для знаходження коефіцієнтів методом Крамера представимо систему в матричному виді:

$$\begin{pmatrix} n & \sum_{i=1}^{n} X_{1i} & \sum_{i=1}^{n} X_{2i} & \ldots & \sum_{i=1}^{n} X_{mi} \\ \sum_{i=1}^{n} X_{1i} & \sum_{i=1}^{n} X_{1i}X_{1i} & \sum_{i=1}^{n} X_{2i}X_{1i} & \ldots & \sum_{i=1}^{n} X_{mi}X_{1i} \\ \sum_{i=1}^{n} X_{2i} & \sum_{i=1}^{n} X_{1i}X_{2i} & \sum_{i=1}^{n} X_{2i}X_{2i} & \ldots & \sum_{i=1}^{n} X_{mi}X_{2i} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \sum_{i=1}^{n} X_{mi} & \sum_{i=1}^{n} X_{1i}X_{mi} & \sum_{i=1}^{n} X_{2i}X_{mi} & \ldots & \sum_{i=1}^{n} X_{mi}X_{mi} \end{pmatrix} \cdot \begin{pmatrix} A_0 \\ A_1 \\ A_2 \\ \ldots \\ A_m \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} Y_iX_{1i} \\ \sum_{i=1}^{n} Y_iX_{2i} \\ \ldots \\ \sum_{i=1}^{n} Y_iX_{mi} \end{pmatrix}$$

Обчислюємо коефіцієнти $A_0, A_1, \ldots, A_m$...

Далі, за аналогією з одномірною моделлю (див. лекцію 3), для кожної крапки обчислюється помилка E_i; потім перебуває сумарна помилка F і значення σ і S з метою визначити, чи приймається висунута гіпотеза про лінійність багатомірного чорного чи ящика ні.

При допомоги підстановок і перепозначень до лінійної множинної моделі приводяться багато нелінійних моделей. Докладно про це розповідається в матеріалі наступної лекції.