УДК 519.85:[512.544.7+519.233.4]
DOI: 10.30838/J.BPSACEA.2312.241023.85.996

СТВОРЕННЯ МОДЕЛІ РЕГРЕСІЇ НА ОСНОВІ АПРОКСИМАЦІЇ ТА ДИСПЕРСІЙНОГО АНАЛІЗУ СТАТИСТИЧНИХ ДАНИХ

ЄРШОВА Н. М., *докт. техн. наук, проф.*

Кафедра комп'ютерних наук, інформаційних технологій та прикладної математики, Придніпровська державна академія будівництва та архітектури, вул. Архітектора Олега Петрова, 24-а, 49005, Дніпро, Україна, тел. +38 (095) 918-01-02, e-mail: nersova107@gmail.com, ORCID ID: 0000-0003-0198-0883

Анотація. ***Постановка проблеми.*** Найскладніше питання у створенні статистичних моделей – це вибір математичної форми зв'язку, тобто аналітичної функції, яка пов'язує елементи системи, що вивчається. Форма рівняння зв'язку встановлюється на основі теоретичних, технологічних міркувань або інтуїції. Коли заздалегідь важко уявити залежність, будують кореляційне поле точок для двох ознак, за розташуванням яких на площині судять про напрям дії і форму зв'язку. Для глибокого і всебічного вивчення статистичних зв'язків використовуються поняття «кореляція» та «регресія». Завдання кореляційного аналізу – встановлення напряму дії та виду форми зв'язку. Завдання регресійного аналізу – побудова математичної моделі регресії у вигляді залежності середнього значення результативної ознаки від факторних ознак. Параметри моделі регресії повинні бути підібрані таким чином, щоб крива, побудована за моделлю, проходила між точками і розташовувалася як можна ближче до всіх точок кореляційного поля, тобто проходила практично через його центр. ***Мета статті*** – створення моделі регресії на основі апроксимації, кореляції та дисперсійного аналізу даних спостережень. ***Результати.*** Виконано апроксимацію даних багатовимірних вибірок активного і пасивного експериментів, отримано апроксимуючі функції і на їх основі моделі регресії в загальному вигляді. На конкретних прикладах установлено зв'язок між факторними ознаками, обрано факторну ознаку, що найбільш значуща за тіснотою зв'язку з результативною ознакою, отримано придатну для прогнозування багатовимірну модель регресії. Виконано багатофакторний дисперсійний аналіз даних активного експерименту на прикладі дослідження впливу на коефіцієнт однорідності бетону марки цементу, типу заповнювача, терміну випробувань та «періоду виготовлення» бетону. Дисперсійний аналіз показує, що найбільш значущі фактори – це марка цементу, термін випробування, «період виготовлення» бетону та незначна їх взаємодія і тип заповнювача. Запропонована методика значно спрощує процес створення моделі регресії. ***Висновки.*** Виконані розрахунки доводять, що: на основі апроксимації і кореляції даних спостережень пасивного експерименту можна встановити зв'язок між факторними ознаками, обрати факторну ознаку, що найбільш значуща за тіснотою зв'язку з результативною ознакою, отримати придатну для прогнозування багатовимірну модель регресії, і це дозволяє спростити процедуру створення моделі регресії шляхом послідовного переходу від простої моделі до складної; дисперсійний аналіз даних активного експерименту дає можливість оцінити вплив окремих факторів і вплив взаємодії поміж ними. Апроксимація доказує, що моделі регресії даних активного експерименту лінійні. Це підтверджує і дисперсійний аналіз, тому що є незначна взаємодія факторів.

Ключові слова: *активний і пасивний експеримент; модель регресії; апроксимація; кореляція; дисперсійний аналіз; результативні і факторні ознаки*

THE CREATION OF A REGRESSION MODEL BASED ON APPROXIMATION AND DISPERSION ANALYSIS OF STATISTICAL DATA

YERSHOVA N.M., *Dr. Sc. (Tech.)., Prof.*

Department of Computer Science, Information Technologies and Applied Mathematics, Prydniprovska State Academy of Civil Engineering and Architecture, 24-a, Architect Oleh Petrov St., Dnipro, 49005, Ukraine, tel. +38 (095) 918-01-02, e-mail: nersova107@gmail.com, ORCID ID: 0000-0003-0198-0883

Abstract. ***Problem statement.*** The most difficult issue when creating statistical models is the choice of a mathematical form of connection, that is, an analytical function that connects the elements of the system being studied. The form of the connection equation is established on the basis of theoretical, technological considerations or intuition. When it is difficult to imagine the dependence in advance, then a correlation field of points is built for two signs, the location of which on the plane determines the direction of action and the form of communication. For a deep and comprehensive study of statistical relationships, the concepts of correlation and regression are used. The task of correlation analysis is to establish the direction of action and the form of communication type. The task of regression analysis is to build a mathematical model of regression in the form of the resulting characteristic average value dependence on factor characteristics. The parameters of the regression model should be selected in such a way that the line constructed by the model passes between the points and is located as close as possible to all points of the correlation field, that is, it passes almost through its center. ***The purpose of the article*** is to create a regression model based on approximation, correlation and dispersion analysis of observational data. ***Results.*** Data approximation for multidimensional samples of active and passive experiments was performed, approximating functions and regression models in general were obtained based on them. On specific examples, the relationship between the factor features was established, the factor feature was selected. The most significant in terms of the connection closeness with the resulting feature, and a multivariate regression model suitable for forecasting was obtained. A multivariate dispersion analysis of the active experiment data was carried out on the research example of the influence on the homogeneity coefficient of concrete cement brand, type of aggregate, test period and “production period: of concrete. Analysis of dispersion shows that the most significant factors are the cement grade, the test period, the “making period” of the concrete and their minor interaction and the type of aggregate. The proposed technique significantly simplifies the process of creating a regression model. ***Conclusions.*** The performed calculations prove that: on the basis of the approximation and correlation of the passive experiment observations, it is possible to establish a relationship between the factor characteristics, to choose the factor characteristic that is the most significant in terms of the connection closeness with the resulting characteristic, to obtain a multivariate regression model suitable for forecasting, and this allows simplify the procedure for creating a regression model by successively moving from a simple model to a complex one; dispersion analysis of active experiment data makes it possible to assess the impact of individual factors and the impact of interaction between them. The approximation proves that the regression models of the active experiment data are linear. This is also confirmed by dispersion analysis, because the interaction of factors is insignificant.

Keywords: *active and passive experiment; regression model; approximation; correlation; dispersion analysis; effective and factor characteristics*

Постановка проблеми. У створенні статистичних моделей найскладніше питання – це вибір математичної форми зв’язку, тобто аналітичної функції, яка пов’язує елементи системи, що вивчається. Форму рівняння зв'язку можна встановити на основі теоретичних, технологічних міркувань або інтуїції, але часто будують кореляційне поле точок для двох ознак, за розташуванням яких на площині судять про напрям дії і форму зв’язку. Для глибокого і всебічного вивчення статистичних зв’язків використовуються поняття «кореляція» та «регресія».

Завдання кореляційного аналізу – це встановлення напряму дії та виду форми зв’язку. Завдання регресійного аналізу – побудова математичної моделі регресії у вигляді залежності середнього значення результативної ознаки від факторних ознак. Параметри моделі регресії повинні бути підібрані таким чином, щоб крива, побудована за моделлю, проходила між точками і розташовувалася якомога ближче до всіх точок кореляційного поля, тобто проходила практично через його центр.

У цій статті пропонується методика створення моделі регресії на основі апроксимації, кореляції та дисперсійного аналізу даних спостережень. Для прискорення процесу впровадження у науку та інженерну практику ефективних комп'ютерних засобів обробки даних експерименту покажемо на прикладах можливості інструментів пакета аналізу Excel. Методика значно спрощує процес створення моделі регресії.

Аналіз останніх досліджень і публікацій. Експерименту належить особливе місце серед способів отримання інформації про внутрішні взаємозв'язки явищ у природі та техніці. Відповідно ускладненню досліджуваних процесів та явищ зростають витрати на апаратуру та проведення експерименту. У ході випробувань збирається велика кількість експериментальних даних, що потребують обробки та аналізу. При цьому тривалість аналізу, осмислення результатів випробувань та їх обліку для коригування характеристик нових виробів дуже значна. Під час проведення спостережень чи експерименту дуже важливо підібрати методи та інструментальні засоби обробки даних експерименту. Інша проблема – створення моделі регресії, яка відповідає даним спостережень. У праці [1] запропонувано методику планування, проведення і обробки даних активного експерименту. Ця робота – є її продовження.

Мета статті − створення моделі регресії на основі апроксимації, кореляції та дисперсійного аналізу даних спостережень.

Результати досліджень. В основу апроксимації і дисперсійного аналізу даних спостережень закладено метод найменших квадратів.

Допустимо, що в результаті експерименту, який складається з N дослідів, отримана таблиця значень одного відгуку y при зміні одного фактора x, тобто функція відгуку задана у вигляді $y = f(x)$.

x	x_1	x_2	…	x_N
y	y_1	y_2	…	y_N

Для отримання теоретичного опису функції відгуку на площині $Y0X$ графічно зображуються таблично задані точки і крива, яка найближче розташована до всіх експериментальних точок.

За формою (зовнішнім виглядом) кривої дають її аналітичний опис у загальному вигляді, тобто у вигляді функції відгуку $\varphi(x)$. У математиці заміна істинної залежності $f(x)$ деякої наближеної $\varphi(x)$, за якої відхилення $\varphi(x)$ від $f(x)$ на аналізованому відрізку було б можливо малим, називається апроксимацією. Функція називається апроксимуючою функцією $\varphi(x)$. Як апроксимуюча функція на практиці зазвичай використовується алгебраїчний багаточлен:

$$\varphi(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n, \qquad (1)$$

коефіцієнти якого $a_0, a_1, a_2, \ldots, a_n$ встановлюються в результаті обробки експериментальних даних. Так, якщо графік апроксимуючої функції пряма, що проходить між експериментальними точками, апроксимуюча функція записується у вигляді:

$$\varphi(x) = a_0 + a_1 x. \qquad (2)$$

Тоді за результатами експерименту потрібно визначити два коефіцієнти: a_0, a_1. Для експериментальних точок, що зображені на рисунку 1, апроксимуючою функцією слід прийняти алгебраїчний багаточлен другого порядку:

$$\varphi(x) = a_0 + a_1 x + a_2 x^2. \qquad (3)$$

І тут потрібно визначити три коефіцієнти: a_0, a_1, a_2. Для двох факторів функція відгуку буде не кривою на площині, а поверхнею в тривимірному просторі, що називається поверхнею відгуку. Поверхня відгуку другого порядку для двох факторів записується у вигляді:

$$\varphi(x) = a_0 + a_1 x_1 + a_2 x_2 + a_{11} x_1^2 + a_{12} x_1 x_2 + a_{22} x_2^2. \qquad (4)$$

І тут потрібно визначити вже шість коефіцієнтів. Розглянемо розв’язання задачі визначення оптимальних значень параметрів a_0, a_1, a_2 алгебраїчного багаточлена (3). Потрібно визначити такі значення параметрів, за яких теоретичні значення апроксимуючої функції, отримані за формулою (3), мали б мінімальні відхилення від фактичних значень функції $y = f(x)$ (рис. 1) для всіх таблично заданих точок. Для цього використовують метод найменших квадратів.

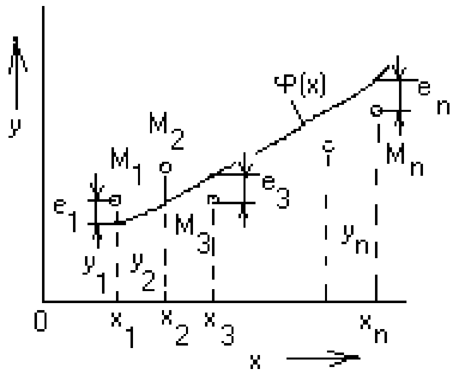

Рис. 1. Геометричний зміст апроксимації

Суть методу полягає в такому:

– записується різниця між значеннями апроксимуючої функції та таблично заданою функцією $y = f(x)$ для кожного таблично заданого x_i: $e_i = \varphi(x_i, a_0, a_1, a_2) - y_i$, що називається відхиленням значення апроксимуючої функції від відповідного табличного значення;

– мінімізують суму квадратів відхилень за всіма таблично заданими x_i тобто $\sum_{i=1}^{n} e_i^2 \to \min$, де n – кількість таблично заданих точок;

– оскільки x_i, y_i відомі (задані таблицею), сума квадратів відхилень є функція параметрів a_0, a_1, a_2. Ця функція позитивна і має мінімум. Позначимо її через $S(a_0, a_1, a_2)$.

Тоді запис методу найменших квадратів має вигляд:

$$S(a_0, a_1, a_2) = \sum_{i=1}^{n} (a_0 + a_1 x_i + a_2 x_i^2 - y_i)^2 \to \min . \quad (5)$$

Маємо функцію багатьох змінних. Необхідна умова мінімуму такої функції полягає в рівності нулю всіх її перших частинних похідних, тобто:

$$\frac{\partial S}{\partial a_0} = 0; \ \frac{\partial S}{\partial a_1} = 0; \ \frac{\partial S}{\partial a_2} = 0. \quad (6)$$

– Точність апроксимації визначається середньоквадратичною похибкою:

$$\varepsilon = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}}. \quad (7)$$

Примітка. Умова мінімуму суми самих відхилень, а не їх квадратів, не вирішує проблеми, оскільки сума $\sum_{i=1}^{n} e_i$ може бути дуже малою і тоді, коли окремі відхилення дуже великі, але мають різні знаки і вони взаємно компенсують одне одного.

Отже, в задачі апроксимації потрібно обчислити такі значення a_0, a_1, a_2, за яких функція $S(a_0, a_1, a_2)$ досягає мінімуму. Це звичайна задача визначення екстремуму функції трьох змінних без обмежень і може бути розв’язана за допомогою надбудови «Пошук рішення» Excel. Найпростіше розв’язати цю задачу шляхом побудови лінії тренду.

Приклад 1. Знайти апроксимуючу функцію по таблично заданих значеннях x_i, y_i, де x_i – обсяг гипсу в суміші, %; y_i – вміст етрингіту, % [1]:

x_i	5	15	25	35	45	55
y_i	5	16	17	16	16,5	15

Технологія підбору апроксимуючої функції шляхом побудови лінії тренду:

1. На аркуші ЕТ ввести два ряди даних, що відповідають один одному, тобто x_i, y_i. Виділити їх.

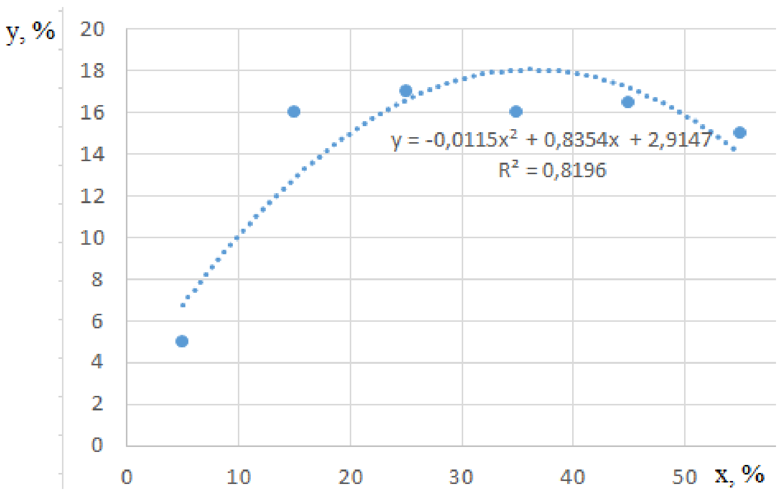

Рис. 2. Точкова діаграма

2. Увійти до пункту меню *Вставка* та вибрати інструмент: ⇒ *Діаграма* ⇒ *Точкова*. У результаті на аркуші з'явиться точкова діаграма. Оформлена діаграма показана на рисунку 2.

3. Спочатку, коли треба, виконаємо формат осі для кращого вигляду діаграми: на осі клацнути правою кнопкою, увійти в контекстне меню і вибрати операцію *Формат осі.* В діалоговому вікні в параметрах осі задати межі (min, max). Потім на будь-якій точці діаграми клацнути правою кнопкою, увійти в контекстне меню і вибрати операцію *Додати лінію тренду.* Відкриється діалогове вікно *Формат лінії тренду* (рис. 3).

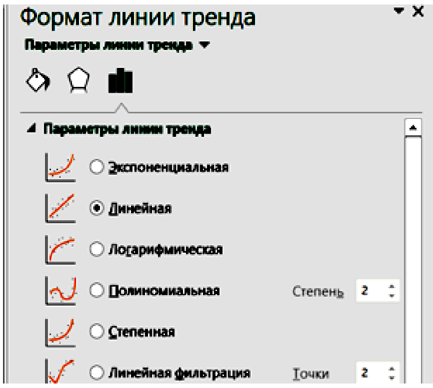

Рис. 3. Діалогове вікно «Формат лінії тренду»

4. У діалоговому вікні *Формат лінії тренду:*

– із запропонованих видів апроксимуючих функцій вибрати поліноміальну лінію тренду 2-го ступеня;

– для отримання на діаграмі аналітичного виразу апроксимуючої функції необхідно в цьому ж вікні встановити галочки, клацнувши в кружках перед *Показувати рівняння на діаграмі* та *Помістити на діаграму величину достовірності апроксимації*. Закрити вікно.

Таким чином, рівняння апроксимуючої функції, отримане шляхом побудови лінії тренду, має вигляд:

$$\varphi(x) = y = 2{,}9147 + 0{,}8354x - 0{,}0115x^2. \quad (8)$$

Достовірність апроксимації R^2 =0,8196.

За апроксимуючою функцією можна написати математичну модель парної нелінійної регресії в загальному вигляді:

$$\overline{Y}_x = a_0 + a_1x + a_2x^2. \quad (9)$$

Математичний запис методу найменших квадратів у цьому випадку має вигляд:

$$S(a_0, a_1, a_2) = \sum_{i=1}^{6}(a_0 + a_1x_i + a_2x_i^2 - y_i)^2 \to \min. \quad (10)$$

Параметри a_0, a_1, a_2 моделі регресії отримаємо за допомогою інструменту *Регресія* пакета аналізу. Вихідна інформація наведена в таблиці 1.

Таблиця 1

Результати роботи інструменту «Регресія»

ВИСНОВОК ПІДСУМКІВ						
Регресійна статистика						
Множінний R	0,9053354					
R-квадрат	0,8196322					
Нормований R-квадрат	0,699387					
Стандартна помилка	2,511047					
Спостереження	6					
Дисперсійний аналіз						
	df	*SS*	*MS*	*F*	*Значення F*	
Регресія	2	85,95892857	42,97946429	6,816341	0,0766017	
Залишок	3	18,91607143	6,305357143			
Разом	5	104,875				
	Коефіцієнти	*Стандартна помилка*	*t-статистика*	*Р-Значення*	*Нижні 95%*	*Верхні 95%*
Y-перетин	2,9147321	3,247134143	0,897632193	0,4355314	-7,419098	13,248562
x	0,8353571	0,253780849	3,291647678	0,046021	0,0277132	1,6430011
x^2	-0,0115179	0,004109665	-2,802626514	0,0677052	-0,024597	0,0015609

Отже, отримали модель парної нелінійної регресії:

$$\overline{Y}_x = 2{,}9147 + 0{,}8354x - 0{,}0115x^2. \quad (11)$$

Коефіцієнт детермінації $R^2 =$ 0,8196, тобто рівняння (8) і (11) однакові.

Отже, за виглядом апроксимуючої функції можна отримати модель парної регресії.

Приклад 2. У таблиці 2 наведено вихідні дані пасивного експерименту. Створити математичну модель регресії на основі апроксимації даних досліджень.

У таблиці 2 позначено: Y_3 – рентабельність, %; X_5 – питома вага робітників у складі промислово-виробничого персоналу (ПВП); X_8 – премії та винагороди на одного працівника, %; X_{17} – невиробничі витрати, %.

Установимо залежність рентабельності від питомої ваги робітників у складі промислово-виробничого персоналу.

1. На робочому аркуші ЕТ вводимо дані таблиці 2.

2. Виділимо вибірки X_5 і Y_3. Спочатку X_5, потім клацнути на Ctrl виділити Y_3.

Таблиця 2

Вихідні дані пасивного експерименту [2]

№	X_5	X_8	X_{17}	Y_3
1	0,78	1,23	17,72	13,26
2	0,75	1,04	18,39	10,16
3	0,68	1,80	26,46	13,72
4	0,70	0,43	22,37	12,85
5	0,62	0,88	28,13	10,63
6	0,76	0,57	17,55	9,12
7	0,73	1,72	21,92	25,83
8	0,71	1,70	19,52	23,39
9	0,69	0,84	23,99	14,68
10	0,73	0,60	21,76	10,05
11	0,68	0,82	25,68	13,99
12	0,74	0,84	18,13	9,68
13	0,66	0,67	25,74	10,03
14	0,72	1,04	21,21	9,13
15	0,68	0,66	22,97	5,37
16	0,77	0,86	16,38	9,86

На рисунку 4 представлена діаграма залежності рентабельності від питомої ваги робітників у складі ПВП.

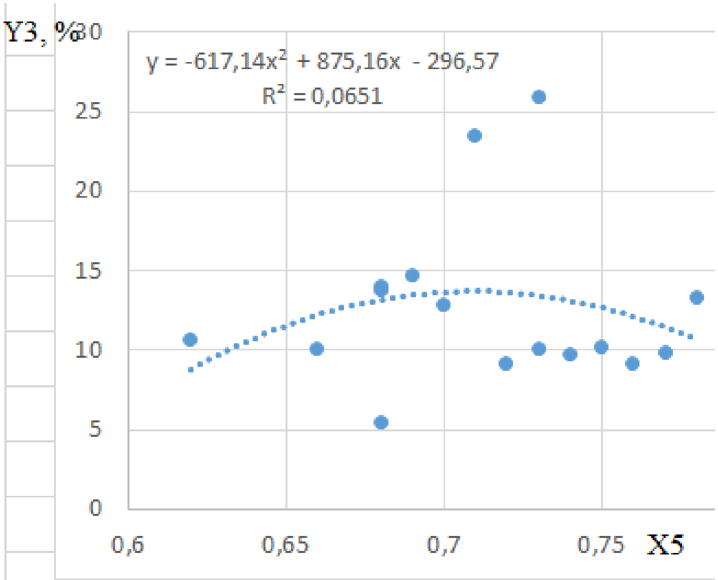

Рис. 4. Залежність рентабельності від питомої ваги робітників у складі ПВП

Отже, рівняння апроксимуючої функції, що отримане шляхом побудови лінії тренду:

$$\varphi(X_5) = y = -296{,}57 + 875{,}16X_5 - 617{,}14X_5^2.$$

Достовірність апроксимації $R^2 = 0{,}0651$.

Залежність рентабельності від премії та винагороди на одного працівника встановлюється аналогічно. Рівняння апроксимуючої функції має вигляд:

$$\varphi(X_8) = y = 13{,}65 - 9{,}6776X_8 + 7{,}9978X_8^2 .$$

Достовірність апроксимації $R^2 = 0{,}5733$.

Рівняння апроксимуючої функції залежності рентабельності від невиробничих витрат:

$$\varphi(X_{17}) = y = -42{,}274 + 5{,}062X_{17} - 0{,}1138X_{17}^2 .$$

Достовірність апроксимації $R^2 = 0{,}0633$.

Аналіз достовірності апроксимації показує, що на рентабельність підприємства суттєво впливають премії та винагороди працівникам.

Установимо кореляційний зв'язок між ознаками за допомогою інструменту *Кореляція* пакета аналізу. Матриця коефіцієнтів парної кореляції наведена в таблиці 3.

Таблиця 3

Матриця коефіцієнтів парної кореляції

	x5	x8	x17	y3
x5	1			
x8	0,2174254	1		
x17	-0,9379766	-0,200748	1	
y3	0,2240482	0,67640838	-0,27235	1

Аналіз коефіцієнтів парної кореляції показує, що між рентабельністю підприємства та преміями і винагородами працівникам спостерігається помітний позитивний кореляційний зв'язок; між питомою вагою робітників у складі промислово-виробничого персоналу та невиробничими витратами – дуже сильний негативний кореляційний зв'язок.

У математичну модель парної регресії слід включити факторну ознаку X_8, що найбільш значуща за тіснотою зв'язку з результативною ознакою. Оскільки факторні ознаки X_5 і X_{17} мають сильний кореляційний зв'язок між собою, їх не можна разом вводити в математичну модель багатовимірної регресії. Тому математична модель багатовимірної лінійної регресії має вигляд:

$$\overline{Y}_x = a_0 + a_1 x_8 + a_2 x_{17}.$$

Отримані апроксимуючі функції нелінійні, тож для прогнозування буде придатна багатовимірна нелінійна модель регресії:

$$\overline{Y}_x = a_0 + a_1 x_8 + a_2 x_{17} + a_{11} x_8^2 + a_{12} x_8 x_{17} + a_{22} x_{17}^2.$$

Приклад 3 [3]. У таблиці 4 наведено умови планування експерименту для дослідження залежності між ознаками: y_1 – міцність на стиснення бетону у віці 28 діб, МПа; x_1 – цементоводне (Ц/В) відношення бетону М200−М400; x_2 – активність цементу $R_ц$, МПа; x_3 – модуль крупності $M_{кр}$; x_4 – зміст домішок $Q_о$, що відмучуються. Матриця планування експерименту наведена в таблиці 5.

Таблиця 4

Умови планування експерименту

Фактор		Рівень фактора			Інтервал
натуральний	кодований	-1	0	1	варіювання
Ц/В	x_1	1,4	2	2,6	0,6
$R_ц$, МПа	x_2	38,8	45,3	51,8	6,5
$M_{кр}$	x_3	1,4	2,2	3	0,8
$Q_о$	x_4	1	3	5	2

Таблиця 5

Матриця планування

№	x_1	x_2	x_3	x_4	y_1	y_2	y_3	$y_{ср}$
1	2,6	51,8	3	5	44,2	43	43,6	43,6
2	2,6	51,8	3	1	49	49,6	47,5	48,7
3	2,6	51,8	1,4	5	42	39,6	41,1	40,9
4	2,6	51,8	1,4	1	45	44	44,2	44,4
5	2,6	38,8	3	5	31,8	32	32,8	32,2
6	2,6	38,8	3	1	35	34	35,4	34,8
7	2,6	38,8	1,4	5	29,6	31	30,6	30,4
8	2,6	38,8	1,4	1	32	33	31,9	32,3
9	1,4	51,8	3	5	20,6	22	20,7	21,1
10	1,4	51,8	3	1	22,5	21	21,9	21,8
11	1,4	51,8	1,4	5	20,8	19,6	18,4	19,6
12	1,4	51,8	1,4	1	21,2	19	20,7	20,3
13	1,4	38,8	3	5	12,9	11	11,8	11,9
14	1,4	38,8	3	1	13,7	13	11,1	12,6
15	1,4	38,8	1,4	5	11	10,4	9,8	10,4
16	1,4	38,8	1,4	1	12	11	10,3	11,1

Діаграма залежності міцності на стиск бетону від Ц/В наведена на рисунку 5.

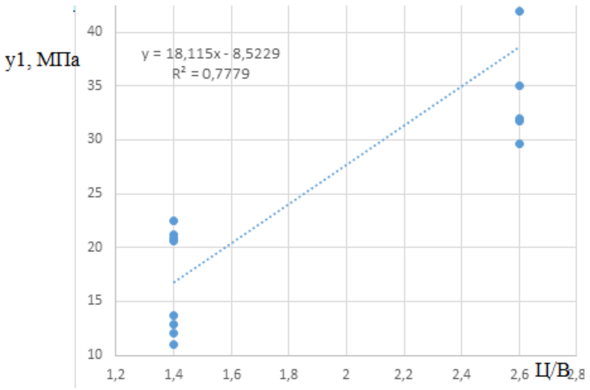

Рис. 5. Залежність міцності на стиснення бетону від Ц/В

Рівняння апроксимуючої функції, отримане шляхом побудови лінії тренду:

$$\varphi(x_1) = y = -8{,}5229 + 18{,}115x_1 .$$

Достовірність апроксимації $R^2 = 0{,}7779$.

Отже, апроксимуючі функції всіх факторів лінійні, тобто і моделі регресії даних активного експерименту лінійні.

У математичну модель регресії активного експерименту включаються всі фактори.

$$\bar{Y}_x = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4,$$

де $a_0, a_1, \ldots, a_4$ – невідомі параметри моделі регресії.

Кореляційна матриця наведена в таблиці 6. Як видно з кореляційної матриці всі коефіцієнти парної кореляції між факторами дорівнюють нулю, тобто фактори незалежні між собою.

Таблиця 6

Матриця коефіцієнтів парної кореляції

	x1	x2	x3	x4	Ycp
x1	1				
x2	-1E-17	1			
x3	0	-4E-17	1		
x4	0	0	0	1	
Ycp	0,8933	0,4246	0,0871	-0,0791	1

Між фактором Ц/В і міцністю бетону на стиск спостерігається сильний кореляційний зв'язок, зв'язок міцності з активністю цементу помітний, а з модулем крупності і вмістом домішок практично відсутній.

Багатофакторний дисперсійний аналіз даних спостережень. У процесі вивчення спільного впливу на результативну ознаку більше двох факторів дисперсійний аналіз ускладнюється. Цей аналіз дає можливість оцінити як вплив окремих факторів, так і вплив взаємодії поміж ними. Повторні випробування можна вважати додатковим фактором.

Приклад 4 [4]. Дослідити вплив на коефіцієнт однорідності бетону K марки цементу R, типу заповнювача A, терміну випробувань T та «періоду виготовлення» бетону N. План експерименту показаний у таблиці 7.

Таблиця 7

Матриця експерименту

Коефіцієнт однорідності бетону					
Марка бетону 200					
Марка цементу 400					
Бетон випробуваний через 1 добу					
			N1	N2	N3
		A1	0,91	0,82	0,6
		A2	0,94	0,76	0,78
		A3	0,89	0,62	0,7
Бетон випробуваний через 28 діб					
		A1	0,9	1,13	1,02
		A2	1,13	0,94	0,94
		A3	0,84	0,85	0,82
Марка цементу 500					
Бетон випробуваний через 1 добу					
		A1	0,78	0,59	0,51
		A2	0,88	0,55	0,55
		A3	0,63	0,46	0,56
Бетон випробуваний через 28 діб					
		A1	0,85	0,72	0,84
		A2	0,64	0,6	0,54
		A3	0,87	0,76	0,62

Заповнювачі: A_1 – гравій річковий; A_2 – щебінь гранітний; A_3 – щебінь вапняковий. Досліджуємо вплив на коефіцієнт однорідності бетону типу заповнювача (фактор A) та «періоду виготовлення» бетону (фактор N). Для цього виконаємо двофакторний дисперсійний аналіз без повторень першої вибірки з таблиці 7.

Зведення дисперсійного аналізу показано у таблиці 8.

Таблиця 8

Дисперсійний аналіз впливу типу заповнювача та «періоду виготовлення» бетону

Дисперсійний аналіз						
Джерело варіації	*SS*	*df*	*MS*	*F*	*P-Значення*	*F критичне*
Рядки	0,038	2	0,019	2,124	0,2351831	6,9442719
Стовпці	0,038	2	0,019	2,098	0,2381534	6,9442719
Похибка	0,036	4	0,009			
Разом	0,113	8				

Аналіз таблиці 8 показує, що розрахункові значення F-критерію цих факторів менші критичних значень. Отже, тип заповнювача та «період виготовлення» не мають суттєвого впливу на коефіцієнт однорідності бетону за випробування зразків через одну добу. Аналогічні результати отримані для другої вибірки таблиці 7.

Досліджуємо вплив на коефіцієнт однорідності бетону марки цементу та терміну випробувань. Для цього виконаємо двофакторний дисперсійний аналіз із повтореннями двох вибірок, одна з яких містить першу та другу вибірки таблиці 7 (марка цементу 400), друга – третю та четверту вибірки (марка цементу 500). Зведення дисперсійного аналізу для марки цементу 400 наведені в таблиці 9.

Таблиця 9

Дисперсійний аналіз впливу марки цементу та терміну випробувань

Дисперсійний аналіз						
Джерело варіації	*SS*	*df*	*MS*	*F*	*P-Значення*	*F критичне*
Рядки	0,1701	1	0,17	11,181	0,00584669	4,7472253
Стовпці	0,0252	2	0,013	0,8291	0,45995296	3,8852938
Взаємодія	0,0161	2	0,008	0,5283	0,60271067	3,8852938
Усередині	0,1826	12	0,015			
Разом	0,3941	17				

Дисперсійний аналіз показує, що для цементу марки 400 суттєво впливає на коефіцієнт однорідності бетону термін випробування $(F > F_{kr})$ і мало впливають «період виготовлення» та їх взаємодія. Для марки цементу 500 спостерігається зворотна картина: термін випробування не суттєво впливає на коефіцієнт однорідності і помітний вплив чинить «період виготовлення» $(F = 4{,}895 > F_{kr} = 3{,}885)$. Також мало впливає їх взаємодія.

Для аналізу впливу на коефіцієнт однорідності бетону марки цементу з таблиці 7 сформовано таблицю, у якій вибірки марок цементу 400 і 500 зібрані разом, а випробування зразків бетону виконувалось через одну добу і 28 діб. Дисперсійний аналіз виявив, що в разі випробування зразків бетону через: добу суттєво впливають на коефіцієнт однорідності бетону марка бетону та «періоди виготовлення» і не впливає їх взаємодія; 28 діб – суттєво впливає на коефіцієнт однорідності бетону марка цементу і не впливають «періоди виготовлення» бетону та їх взаємодія. Зведення дисперсійного аналізу всіх вибірок таблиці 7 представлене в таблиці 10.

Додатково обчислено середні значення коефіцієнта однорідності бетону. Дисперсійний аналіз показує, що найбільш значущими факторами є марка цементу, термін випробування, «період виготовлення» бетону та незначна їх взаємодія. Середнє значення коефіцієнта однорідності бетону зменшується зі зміною марки цементу: K_{400} = 0,757 та K_{500} = 0,612 за випробування через добу; K_{400} = 0,952 і K_{500} = 0,715 під час випробування через 28 діб, тобто зі збільшенням віку бетону коефіцієнт однорідності збільшується, що пояснюється стабілізацією структуро-утворення та зняттям термічної напруги. Позначимо: коефіцієнт однорідності бетону $K = x_1$, марка цементу $R = x_2$, термін випробувань $T = x_3$, «період виготовлення» бетону $N = x_4$.

Таблиця 10

Дисперсійний аналіз всіх вибірок

ВИСНОВОК	N1	N2	N3	Разом		
Счет	3	3	3	9		
Сума	2,54	2,2	2,08	6,82		
Середнє	0,847	0,733	0,693	0,7578	Кср	
Дисперсія	0,019	0,011	0,008	0,0141		
A1						
Счет	3	3	3	9		
Сума	2,87	2,92	2,78	8,57		
Середнє	0,957	0,973	0,927	0,9522	Кср	
Дисперсія	0,023	0,02	0,01	0,0139		
A1						
Счет	3	3	3	9		
Сума	2,29	1,6	1,62	5,51		
Середнє	0,763	0,533	0,54	0,6122	Кср	
Дисперсія	0,016	0,004	7E-04	0,0181		
A1						
Счет	3	3	3	9		
Сума	2,36	2,08	2	6,44		
Середнє	0,787	0,693	0,667	0,7156	Кср	
Дисперсія	0,016	0,007	0,024	0,0148		
Разом						
Счет	12	12	12			
Сума	10,06	8,8	8,48			
Середнє	0,838	0,733	0,707			
Дисперсія	0,02	0,035	0,029			
Дисперсійний аналіз						
Джерело варіації	*SS*	*df*	*MS*	*F*	*P-Значення*	*F критичне*
Вибірка	0,547	3	0,182	13,71	2,052E-05	3,0087866
Стовпці	0,116	2	0,058	4,3727	0,0240292	3,4028261
Взаємодія	0,052	6	0,009	0,6476	0,6915676	2,5081888
Усередині	0,319	24	0,013			
Разом	1,034	35				

Загальний вигляд моделі багатовимірної лінійної регресії коефіцієнта однорідності бетону:

$$\bar{Y}_x = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4,$$

де $a_0, a_1, ..., a_4$ – невідомі параметри моделі регресії.

Отже, апроксимація, кореляція та дисперсійний аналіз даних спостережень позволяють створювати математичні моделі регресії дослідних ознак.

Висновки

Виконані розрахунки показують:

1. На основі апроксимації і кореляції даних спостережень пасивного експерименту можна встановити зв'язок між факторними ознаками, обрати факторну ознаку, найбільш значущу за тіснотою зв'язку з результативною ознакою, отримати придатну для прогнозування багатовимірну модель регресії. Це дозволяє спростити процедуру створення моделі регресії шляхом послідовного переходу від простої моделі до складної.

2. Дисперсійний аналіз даних активного експерименту дає можливість оцінити вплив окремих факторів і вплив взаємодії поміж ними. Апроксимація доводить, що моделі регресії лінійні. Це підтверджує і дисперсійний аналіз, тому що є незначна взаємодія факторів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Ершова Н. М., Деревянко В. Н., Тимченко Р. А., Шаповалова О. В. Обработка данных средствами Excel при планировании эксперимента : учеб. пособ. для вузов. Днепропетровск : ПГАСА, 2012. 350 с.

2. Дубров А. М., Мхитарян В. С., Трошин Л. И. Многомерные статистические методы для экономистов и менеджеров : учеб. Москва : Финансы и статистика, 2000. 352 с.

3. Дворкін Л. Й. Експериментально-статистичне моделювання при проектуванні складів бетонів : навч. посіб. Київ : Видавничий дім «Кондор», 2020. 205 с.

4. Вознесенский В. Статистические решения в технологических задачах. Кишинев : Картя Молдовеняскэ, 1969. 232 с.

5. Митропольский А. К. Техника статистических вычислений. Москва : Главная редакция физико-математической литературы издательства «Наука», 1971. 576 с.

REFERENCES

1. Yershova N.M., Derevyanko V.N., Timchenko R.A. and Shapovalova O.V. *Obrabotka dannykh sredstvami Excel pri planirovanii eksperimenta: ucheb. posobiye dlya vuzov* [Data processing using Excel when planning an experiment : textbook manual for universities]. Dnipropetrovsk : PSACEA Publ., 2012, 350 p. (in Russian).

2. Dubrov A.M., Mkhitaryan V.S. and Troshin L.I. *Mnogomernyye statisticheskiye metody dlya ekonomistov i menedzherov : uchebnik* [Multidimensional statistical methods for economists and managers : textbook]. Moscow : Finance and Statistics Publ., 2000, 352 p. (in Russian).

3. Dvorkin L.Yo. *Eksperymental'no-statystychne modelyuvannya pry proektuvanni skladiv betoniv : navch. posib*. [Experimental-statistical modeling in the design of concrete compositions : training manual]. Kyiv : “Condor” Publishing House, 2020, 205 p.(in Ukrainian).

4. Voznesenskiy V. *Statisticheskiye resheniya v tekhnologicheskikh zadachakh* [Statistical solutions in technological problems]. Kishinev : Kartya Moldovenyaske Publ., 1969, 232 p. (in Russian).

5. Mitropol'skiy A.K. *Tekhnika statisticheskikh vychisleniy* [Statistical computing technique]. Moscow : Main editorial office of physical and mathematical literature of the “Science” Publishing House, 1971, 576 p. (in Russian).

Надійшла до редакції: 11.09.2023.