

Переваги багатофакторного регресійного аналізу перед ANOVA

А.В. Бондаренко, С.М. Лапач

КПІ ім. Ігоря Сікорського, Київ, Україна

Анотація. При моделюванні технологічних процесів необхідно враховувати велику кількість факторів. До цього часу в учбових курсах і в практичній діяльності часто застосовується дисперсійний аналіз (ANOVA), хоча вже досить давно показана можливість використання замість нього регресійного аналізу, особливо при великій кількості факторів. В роботі детально, на прикладі, розглядаються переваги регресійного аналізу над ANOVA при моделюванні технологічних процесів. За приклад взято моделювання виготовлення болтів зі скловолокніту з урахуванням трьох факторів. Приведені моделі регресійного аналізу і ANOVA, їх статистичні характеристики. Виконано детальний порівняльний аналіз їх застосування для моделювання і оптимізації технологічних процесів. Показані значні переваги регресійного аналізу порівняно з дисперсійним для моделювання технологічних процесів.

Ключові слова: регресійний аналіз; ANOVA; моделювання технологічних процесів

За ілюстративний приклад взята задача зі Степнова М.М. по механічним випробуванням. Досліджувався вплив технологічних параметрів при виготовленні болтів зі скловолокніту на межу міцності. Виробництво болтів (діаметр 14 мм) зі скловолокніту виконувалось пресуванням. Вивчався вплив таких факторів як температура при пресуванні, час витримки в прес формі і тиск при пресуванні (див. табл. 1, 2).

Таблиця 1.

Опис факторів

Назва і натуральне позначення фактору	Умовне позначення фактору	Рівні варіювання фактору			
		0	1	2	3
Температура при пресуванні, t°C	X ₁	135	145	155	165
Час витримки в прес формі, τ(хв.)	X ₂	1	3	6	–
Тиск при пресуванні P (МПа)	X ₃	20	30	40	50

Таблиця 2.

Середні значення межі міцності болтів, МПа (t – час витримки в хв.)

Тиск, МПа при пресуванні	Температура при пресуванні											
	T=135°C			T=145°C			T=155°C			T=165°C		
	t=1	t=3	t=6	t=1	t=3	t=6	t=1	t=3	t=6	t=1	t=3	t=6
20	106,3	96,8	83,3	79,8	69,6	76	87,6	75,62	77,5	72,3	67,4	53,2
30	131,3	82,8	78	90,5	79,4	93,4	99,2	81,76	82,7	49,68	75,1	73,7
40	96,2	98,4	92	79,8	73,4	66,4	78,7	86,16	80,8	76,82	76,2	80,7
50	92,4	86,8	91,3	97,0	65,8	83,6	69,8	74,5	70,2	65,7	54,1	64,6

Задача розв'язувалась з використанням три факторного дисперсійного аналізу. Аналізувались головні ефекти і всі взаємодії (подвійні і потрійні). Всі ефекти, що піддавалися аналізу були лінійні, виходячи з задачі дисперсійного аналізу. Результати дисперсійного аналізу приведені в табл. 3.

Таблиця 3.

Результати дисперсійного аналізу

Компонент дисперсії	Сума квадратів, Q	Число степенів свободи, ν	Дисперсія, S^2	Розрахункове значення критерію Фішера, F	Критичне значення критерію Фішера F при $\alpha=0,05$
Між температурами пресування, X_1	22373	3	7458	21,43	2,65
Між часом витримки, X_2	3190	2	1595	4,59	3,04
Між тиском пресування, X_3	2674	3	891	2,56	2,65
Взаємодія між температурою і часом, X_1X_2	3787	6	631	2,03	2,14
Взаємодія між температурою і тиском, X_1X_3	4551	9	506	1,63	1,92
Взаємодія між часом і тиском, X_2X_3	1909	6	318	1,02	2,14
Взаємодія між температурою, часом і тиском, $X_1X_2X_3$	10279	18	571	1,84	1,66
Залишкова	59792	192	311	–	–
Повна (загальна)	108555	239	454	–	–

За результатами дисперсійного аналізу сформульовані наступні висновки: “Оптимальними режимами є: температура при пресуванні 135°C , час витримки в прес формі 1 хв., тиск при пресуванні 20МПа. В процесі виробництва більшу увагу слід приділяти контролю температури пресування і часу витримки і меншу – контролю тиску”.

З таблиці дисперсійного аналізу випливає, що статистично значимим є тільки температура, тиск і потрійна взаємодія всіх факторів.

Скажемо так: для таких висновків власне дисперсійного аналізу і не потрібно. Оптимізацію можна провести просто за табл. 2. Про силу впливу, звичайно, можна подивитись по сумі квадратів в дисперсійному аналізі, але досвідчений дослідник може зробити ці висновки просто з тієї ж табл. 2. Тобто, власне дисперсійний аналіз практично не додає нам інформації про процес, крім визначення статистично значимих факторів. Ніякого обчислювального експерименту за моделлю дисперсійного аналізу звісно робити неможливо.

Використання регресійного аналізу замість дисперсійного обґрунтувалось в [1, 2, 3]. Зауважимо, що і використання нечислових (таких, які вимірюються за номінальною шкалою) факторів в моделях регресійного аналізу теж обґрунтовано майже півстоліття тому.

Рівні варіювання факторів описані в табл. 1 Виходячи з кількості факторів (три і чотири), зрозуміло, що, на відміну від дисперсійного аналізу, можлива побудова моделі більш складної, ніж лінійна. Можливе врахування головних ефектів до другої і третьої степенів відповідно. При цьому і ефекти взаємодій теж можуть мати високу степінь, що дозволяє описувати складні форми поверхні відгуку. При побудові нижче приведеної моделі аналізувалися головні ефекти всіх факторів до третього степені включно з усіма утвореними ними ефектами подвійних і потрійних взаємодій.

За вихідними даними ми маємо план повного факторного експерименту (ПФЕ) виду $3^1 \times 4^2 // 48$. ПФЕ є ідеальним з точки зору стійкості результатів і можливостей побудови регресійної моделі. З використанням програмного засобу ПРІАМ (планування, регресія і аналіз моделі) [4] отримана модель

$$Y = 80,3738 - 12,0258x_1 - 4,68525u_1 - 4,06816x_2 - 3,08958z_3 + 5,19068x_1x_2 - 3,74875z_1u_3 - 6,33191x_1z_2u_3 - 4,27026x_1x_2z_3 + 5,57012x_1x_2u_3 + 5,56765u_1x_2u_3 + 5,69433u_1z_2x_3 + 3,14227z_2 + 3,55258x_2x_3 + 3,73194z_2u_3 - 3,56385x_1u_3 + 3,51315u_1x_3 - 3,44359x_1z_2 + 2,62121z_1x_2x_3 - 2,43921u_1x_2z_3 - 3,09988u_1x_2x_3 + 2,47253x_1z_2z_3 + 2,42615u_1z_2z_3 - 1,69875x_1z_3 + 2,08005u_1u_3 + 1,94753u_1z_2 - 1,27025x_3 - 1,20641z_1z_2 + 1,39374u_1x_2 - 1,80344x_1x_2x_3 + 1,85998x_1z_2x_3 + 0,95875u_1z_3 + 0,90825u_3 + 0,672083z_1 + 1,2376z_1z_2x_3 + 0,907073z_2z_3 - 1,04047z_1x_2u_3$$

де:

$$x_1 = 0,0666667*(X_1 - 150); \quad z_1 = 2,25*(x_1^2 - 0,555556); \quad u_1 = 3,75*(x_1^3 - 0,911111*x_1);$$

$$x_2 = 0,375*(X_2 - 3,33333); \quad z_2 = 1,80148*(x_2^2 - 0,184211*x_2 - 0,59375);$$

$$x_3 = 0,0666667*(X_3 - 35); \quad z_3 = 2,25*(x_3^2 - 0,555556); \quad u_3 = 3,75*(x_3^3 - 0,911111*x_3)$$

Модель має високі статистичні характеристики. При хорошій інформативності (параметр $\gamma = 3$) вона пояснює 99,18% розсіяння (табл. 4.). Модель адекватна і має ідеальні показники структурної і обчислювальної стійкості.

Таблиця 4.

Статистичні характеристики моделі

Характеристика	Позначення	Значення
Множинний коефіцієнт кореляції	R	0,9959
Скорегований множинний коефіцієнт кореляції	$R_{кор}$	0,9823
Розрахункове значення критерію Фішера для R	F_R	36,916
Критичне значення для критерію Фішера	F_{α, v_1, v_2}	2,544
Степені свободи	v_1	36
	v_2	11
Частка розсіювання, пояснювана моделлю	R^2	99,18%
Критерій Бокса і Веца	γ	3
Дисперсія залишкова	$S_{зали}^2$	7,256
Степені свободи	v_1	36
	v_2	11
Розрахункове значення критерію Фішера	$F_{ад}$	28,510
Критичне значення для критерію Фішера	F_{α, v_1, v_2}	2,544
Рівень значущості	α	0,05
Структурна стійкість	$\text{Max}(r_{ij})$	0
Обчислювальна стійкість	cond	1

Точність апроксимації (див. табл. 5) висока і рівномірна: середні значення точності апроксимації, як абсолютної, так і відносної, майже не відрізняються від максимальних відхилень.

Таблиця 5.

Точність опису

Середня абсолютна похибка апроксимації	1,065
Середня похибка апроксимації (%)	1,395
Максимальна абсолютна похибка апроксимації	9,07
Максимальна похибка апроксимації	9,37

З табл. 6 видно, що всі складові моделі, за винятком X_1 і взаємодії X_1X_2 мають домінуючу нелінійну складову. Це означає, що поверхня відгуку регресійної моделі має складну форму. Крім того, другим за силою впливу ефектом є потрійна взаємодія $X_1X_2X_3$.

Тобто, не враховувати всі фактори без зниження якості аналізу, неможливо.

Наявність адекватної і інформативної моделі дозволяє виконувати обчислювальний експеримент як з метою дослідження процесу, так і його оптимізації. Можливо проводити оптимізацію, прогноз, будувати графіки маргінальних рівнянь регресії і поверхні відгуку, що дозволяє отримати інформацію про процес як в числовій, так і в наочній графічній формі.

Таблиця 6.

Сила впливу факторів

№п/п	Фактор	Частка впливу (%)	В тому числі нелінійна складова (%)
1	Температура при пресуванні, t°C	45,91	6,24
2	Час витримка в прес формі, τ (хв.)	7,32	6,83
3	P (МПа)	5,37	4,93
4	Взаємодія температури t°C і часу τ (хв.)	7,25	2,86
5	Взаємодія температури t°C і тиску P(МПа)	9,41	9,41
6	Взаємодія часу τ (хв.) і тиску P(МПа)	4,21	4
7	Взаємодія температури t°C, часу τ (хв.) і тиску P(МПа)	19,77	16,83

Висновки

Можливо виділити наступні переваги регресійного аналізу порівняно с дисперсійним при дослідженні і оптимізації технологічних процесів.

1. Отримується не тільки якісна оцінка виду «впливає/не впливає», а також можливість кількісної оцінки сили впливу окремих ефектів.

2. Можливість врахування нелінійностей і побудови моделей складніших і точніших за лінійчаті.

3. Можливо оцінити якість моделі за конкретними числовими характеристиками і відповідно можливі межі її застосування, надійність і обґрунтованість висновків за нею.

4. Можливість обчислювального експерименту за отриманою моделлю для дослідження і оптимізації процесу.

5. Наявність програмних засобів, зручних для використання проти незручностей практично ручного розрахунку в дисперсійному аналізі при числі факторів три і більше.

Таким чином, використання регресійного аналізу замість дисперсійного має ряд переваг як наукового так і прикладного характеру.

Список літератури

- С.М. Лапач Розширення сфери застосування теорії планування експериментів // Матеріали VI Міжнародної науково-практичної конференції «Математика в сучасному університеті», Київ, 28-29 грудня 2017р. С.76–79.
- А.Ю. Долішній, С.М. Лапач Застосування регресійного аналізу замість ANOVA / Тези доповідей загально-університетської науково-технічної конференції молодих вчених та студентів, присвяченої дню Науки. Секція "Машинобудування", підсекція "Технологія машинобудування" – К: НТУУ "КПІ", 2013. С. 31–32.
- Norman R. Draper, Harry Smith Applied regression Analysis. Third edition / New York: Wiley-Interscience, 1998. 736 p.
- Теорія планування експериментів: Виконання розрахунково-графічної роботи [Електронний ресурс]: навч. посіб. для студ. спеціальності 131 «Прикладна механіка», спеціалізації «Технологія машинобудування»/ С.М. Лапач ; КПІ ім. Ігоря Сікорського. – Електронні текстові данні (1 файл: 3,31 Мбайт). – Київ: КПІ ім. Ігоря Сікорського, 2020. – 86 с. <https://ela.kpi.ua/handle/123456789/38858>

Advantages of multivariate regression analysis over ANOVA

A.V. Bondarenko, S.M. Lapach

Abstract. When modeling technological processes nowadays, it is necessary to take into account a large number of factors. Until now, analysis of variance (ANOVA) is often used in educational courses and in practical activities, although the possibility of using regression analysis instead, especially with a large number of factors, has been shown for quite some time. The paper examines in detail, using an example, the advantages of regression analysis over ANOVA in the modeling of technological processes. As an example, modeling of the manufacture of fiberglass bolts taking into account three factors is taken. Models of regression analysis and ANOVA, their statistical characteristics are given. A detailed comparative analysis of their application for modeling and optimization of technological processes was performed. Significant advantages of regression analysis compared to dispersion analysis for modeling technological processes are shown.

Keywords: regression analysis; ANOVA; modeling of technological processes