УДК 651.92:001.817

Р.О. Бакарджиєв, доц., канд. техн. наук
Таврійський агротехнологічний університет, zcmt.nncimesg@gmail.com
А.О. Комаров
Київський національний університет ім. Т. Шевченка

Попередня оцінка і обробка даних при регресійному аналізі

Представлено методи попередньої оцінки і обробки даних активного експерименту з застосуванням статистичних функцій Excel і пакетом прикладних програм Statistica при проведенні регресійного аналізу.

На конкретних прикладах проілюстровано перевірку вибірки на нормальність статистичного розподілу, оцінку сумнівних результатів за критерієм Стюдента, перевірку відтворюваності результатів за критерієм Кохрена. Аналіз виконувався без залучення довідкових таблиць зі значеннями розподілів оціночних критеріїв, визначені кореляційні відношення залежних і незалежних ознак.

Наведені способи дають змогу швидко з високою точністю оцінити вибірку з будь–якою кількості факторів на придатність для регресійного аналізу.

регресійний аналіз, нормальність статистичного розподілу, критерій Стюдента, критерій Кохрена, кореляційне відношення

Р.А. Бакарджиев, доц., канд. техн. наук
Таврический агротехнологический университет;
А.А. Комаров
Киевский национальный университет им. Т. Шевченко

Предварительная оценка и обработка данных при регрессионном анализе

Представлены методы предварительной оценки и обработки данных активного эксперимента с применением статистических функций Excel и пакетом прикладных программ Statistica при проведении регрессионного анализа.

На конкретных примерах приводится проверка выборки на нормальность статистического распределения, оценка сомнительных результатов по критерию Стюдента, проверка воспроизводимости результатов по критерию Кохрена. Анализ выполняется без использования справочных таблиц со значениями распределений оценочных критериев, определены взаимные корреляционные отношения зависимых и независимых признаков.

Приведенные способы позволяют быстро с высокой точностью оценить выборку с любым количеством факторов на пригодность для регрессионного анализа.

регрессионный анализ, нормальность статистического распределения, критерий Стюдента, критерий Кохрена, корреляционное отношение

Постановка проблеми. Регресійний аналіз – статистичний метод дослідження впливу на залежну змінну однієї або декількох незалежних змінних. Його метою є пошук таких комбінацій незалежних ознак регресійної експериментально-статистичної моделі, які з найбільшою статистичною достовірністю прогнозують значення залежної ознаки [1].

Умовами застосування методу регресійного аналізу є:

– між кожною з незалежних ознак $\boldsymbol{x}$ і залежною $\boldsymbol{y}$ повинні бути причинно–наслідкові залежності;

– число об’єктів дослідження має бути у кілька разів більше числа прогностичних (пояснюючих) ознак, тобто обсяг вибірки повинен в 3–5 разів перевищувати кількість факторів.

– усі аналізовані ознаки, як незалежні, так і залежна, повинні бути кількісними;

– повторення залежних ознак повиннні мати нормальний розподіл з рівними дисперсіями;

– взаємозв'язки між кожною з незалежних ознак ***x*** і залежною ***y*** в інтервалі досліджуваних значень описується функцією одного показника;

– кожне значення ***y*** незалежне від іншого для кожного значення x_i;

– необхідна відсутність взаємної кореляції незалежних ознак. Якщо будь–які з незалежних ознак сильно або середньо корельовані між собою, то необхідно залишити для регресійного аналізу ті з них, які мають більш сильну кореляцію із залежною ознакою [2].

Аналіз останніх досліджень і публікацій. Одні з цих вимог забезпечуються організацією експерименту [1], виконання інших, зокрема, оцінка отриманої функції відгуку, аналізується програмами статистичної обробки [3], проте виконання деяких з них повинно перевірятися перед статистичною обробкою даних самим експериментатором [2].

Проте, як показує аналіз наукових робіт останніх років, попередня оцінка даних перед регресійним аналізом у кращому разі обмежується лише виявленням недостовірних даних.

Постановка завдання. Метою статті є представлення методів проведення попередньої оцінки і обробки отриманого масиву експериментальних даних при регресійному аналізі активного експерименту. Обробка виконується з використанням пакету прикладних програм Statistica і статистичних функцій табличного процесора Microsoft Excel, що виключає застосування довідкових таблиць з показниками розподілів, чим суттєво підвищує швидкість і точність розрахунків.

Виклад основного матеріалу. Для ілюстрації процесів перевірок вимог до застосування методу регресійного аналізу нами взятий приклад залежності щільності **r** (кг/м^3) паливних брикетів від довжини часток соломи ***l*** (мм), умісту **d** (%) зв'язуючої речовини і кута **a** (град) конусності матриці (табл. 1), отриманої за трирівневою матрицею повнофакторного плану другого порядку для трьох факторів [4]. Дослідження виконані з трикратною повторністю — мінімальною кількістю, яка забезпечує 95 % надійність досліду [5].

Таблиця 1 – Результати досліджень щільності паливних брикетів з соломи

№ п.п.	Фактори			Щільність брикету ρ, кг/м³			
	Довжина часток *l*, мм	Уміст зв'язуючого δ, %	Конусність матриці α, град	Повторності			Середнє $Y_{ср}$
				Y_1	Y_2	Y_3	
1	20	0.0	2	596.5	788.9	672.6	686
2	40	0.0	2	740.5	489.4	576.1	602
3	20	9.0	2	516.4	768.6	605.0	630
4	40	9.0	2	683.2	444.4	525.4	551
5	20	0.0	6	559.3	846.2	658.4	688
6	40	0.0	6	734.5	501.7	584.9	607
7	20	9.0	6	510.5	784.9	603.6	633
8	40	9.0	6	674.0	460.3	536.7	557
9	20	4.5	4	527.0	784.5	617.5	643
10	40	4.5	4	692.5	457.7	538.8	563
11	30	0.0	4	501.6	771.3	593.1	622
12	30	9.0	4	718.2	452.4	539.4	570
13	30	4.5	2	473.2	763.3	566.5	601
14	30	4.5	6	774.4	472.7	567.9	605

Перевірка вибірки на нормальність розподілу виконується згідно міжнародного стандарту ІСО 3479–97 за критерієм Шапіро–Уїлка. Її краще здійснювати побудовою гістограм пакетом прикладних програм Statistica.

Результати оцінки подані в табл. 2.

З неї бачимо, що для всіх повторень виконується умова $p > a$, де a – прийнятий рівень статистичної значущості, a = 0.05. Таким чином всі повторення мають нормальний статистичний розподіл, тому для їх оцінок можуть бути застосовані параметричні критерії Фішера, Стьюдента, Кохрена та інші.

Таблиця 2 – Відповідність повторень нормальному розподілу

Повторність	Критерій Шапіро-Уилка ***W***	Рівень значущості ***p***
Y_1	0.89884	0.1085
Y_2	0.86169	0.07467
Y_3	0.93937	0.41033

Розглянемо оцінку сумнівних і виключення помилкових значень (промахів) для такої повторності, коли у вибірці є тільки одне аномальне значення із трьох.

Найбільш уживаним для цього є критерій Стьюдента [5], значення якого табульовані для вибірок обсягом від двох до ∞. Так як одні автори рекомендують проводити обчислення за всією вибіркою, інші – з виключенням сумнівних результатів, з огляду на лише трикратне повторення, слід проводити обчислення за всією вибіркою. У першу чергу сумнівними є мінімальні і максимальні значення, які знаходяться відповідними статистичними функціями MS Excel МИН і МАКС. При розрахунках за формулою (1) для них знаходять $t_ф$ – фактичні значення критерію Стьюдента (колонки F і H), які порівнюється з $t_{a(n)_т}$ – з його табличними критичними значеннями (чарунка E17), знайденим за статистичною функцією Excel СТЬЮДЕНТ.ОБР.2Х при відповідному рівні значущості a і ступенях вільності $n = n - 1$.

$$t_ф = \frac{x_i - \overline{x}}{s}. \quad (1)$$

де $\overline{x}$ – середнє арифметичне значення, яке отримується за допомогою статистичної функції MS Excel СРЗНАЧ;

дs – вибірковий середній квадратичний відхил, який визначається функцією СТАНДОТКЛОН.В.

Таблиця 3 – Результати оцінки сумнівних даних за ***t***-критерієм Стьюдента

	A	B	C	D	E	F	H
1	№ п.п.	Y_1	Y_2	Y_3	s	$t_{(max)}$	$t_{(min)}$
2	1	596.5	788.9	672.6	96.89	0.924	1.062
3	2	740.5	489.4	576.1	127.50	0.883	1.086
4	3	516.4	768.6	605.0	127.95	0.888	1.083
5	4	683.2	444.4	525.4	121.48	0.878	1.089
6	5	559.3	846.2	658.4	145.72	0.883	1.086
7	6	734.5	501.7	584.9	117.97	0.893	1.080
8	7	510.5	784.9	603.6	139.56	0.878	1.089
9	8	674.0	460.3	536.7	108.26	0.893	1.080
10	9	527.0	784.5	617.5	130.59	0.888	1.083
11	10	692.5	457.7	538.8	119.24	0.883	1.086
12	11	501.6	771.3	593.1	137.14	0.878	1.089
13	12	718.2	452.4	539.4	135.52	0.868	1.094
14	13	473.2	763.3	566.5	148.07	0.863	1.096
15	14	774.4	472.7	567.9	154.25	0.858	1.098
16		Рівень значущості α			0.05		
17		Табличне значення $t_т$			4.3027		

E2=СТАНДОТКЛОН.В(B2:D2)
F2=(СРЗНАЧ(B2:D2)-МИН(B2:D2))/E2
G2=(СРЗНАЧ(B2:D2)-МИН(B2:D2))/E2
E17=СТЬЮДЕНТ.ОБР.2Х(E16;СЧЁТ(B2:D2)-1)

Дані, для яких виконується умова $t_ф \ ^3 \ t_{a(n)_т}$, виключаються із розгляду.

Застосування цього методу розглянуто у табл. 3. За її результатами всі мінімальні і максимальні значення повторень є достовірними з прийнятим рівнем значущості.

Перевірка належності всіх вибірок до однієї генеральної сукупності здійснюється оцінкою статистичної значущості різниці між парами середніх арифметичних вибірок з використанням критерію Стьюдента чи за порівнянням їхніх довірчих інтервалів. На наш погляд доцільно використовувати останній спосіб, здійснюючи побудову блокових діаграм з графічним указанням середнього, його похибки і довірчого інтервалу для середнього арифметичного, яке виконується у ППП Statistica (рис. 1).

З аналізу рис. 1 бачимо, що всі три вибірки мають загальний довірчий інтервал середнього, який знаходиться у межах від *A* до *B*, що свідчить про їх належність до однієї генеральної сукупності.

Для перевірки гіпотези про однорідність (належність до однієї генеральної сукупності) дисперсій всіх вибірок використовується критерій Кохрена *G* (*G*–критерій).

Цей критерій використовується при активному експерименті з однаковою повторністю (більше двох) *n*. Він залежить від числа дослідів *n* та кількості варіантів (повторення) досліду *k* і представляє собою перевірку однорідності дисперсій обчисленням частки від ділення максимальної дисперсії $s^2_{\max}$ на суму всіх дисперсій $\sum_i^k s_i^2$

$$G = \frac{s^2_{\max}}{\sum_i^k s_i^2}. \qquad (2)$$

Отримана частка порівнюється із критичним (табличним) значенням $G_{a(k,n)_т}$. Хоч значень функції розподілу Кохрена немає ні в MS Excel, ні пакеті прикладних програм Statistica, проте їх можна апроксимувати *F*–розподілом Фішера [3].

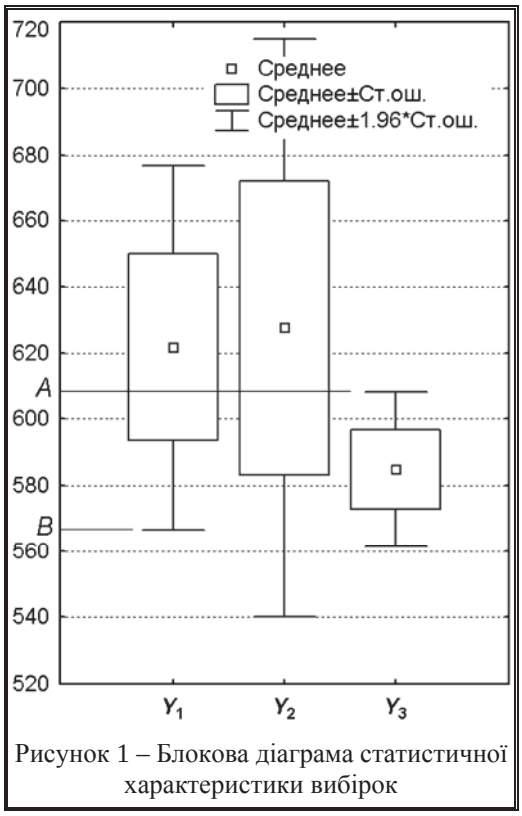

Рисунок 1 – Блокова діаграма статистичної характеристики вибірок

	A	B	C	D
1	Y1	Y2	Y3	s²
2	596.5	788.9	672.6	6258.30
3	740.5	489.4	576.1	10837.77
4	516.4	768.6	605.0	10913.69
5	682.2	444.4	525.4	9838.57
6				46
7				68
8				83
9				02
10				75
11				03
12	501.6	771.3	593.1	12537.46
13	718.2	452.4	539.4	12244.23
14	473.2	763.3	566.5	14615.76
15	774.4	472.7	567.9	15861.46
16	Вільність повторності дослідів **ν**			2
17	Повторення дослідів ***k***			14
18	Рівень значущості α			0.05
19	Розрахунковий критерій Кохрена ***G***			0.1003
20	Критерій Фішера ***F***			7.0533
21	Критичний критерій Кохрена $G_т$			0.3517

D2=ДИСПР(A2:C2)
D15=ДИСПР(A15:C15)
D16=СЧЁТ(A15:C15)-1
D17=СЧЁТ(C2:C15)
D19=МАКС(D2:D15)/СУММ(D2:D15)
D20=F.ОБР.ПХ(D18/D17;D16;(D17-1)*(D16))
D21=D20/(D20+D17-1)

У нашому разі це виконується з використанням статистичної функції Excel F.ОБР.ПХ.

$$G_{(k,\nu)_т} = \frac{F_{\alpha/k,\nu,(k-1)\cdot\nu}}{F_{\alpha/k,\nu,(k-1)\cdot\nu} + k - 1}, \qquad (3)$$

Якщо $G < G_{a(k,n)_т}$, то нульова гіпотеза підтверджується — всі вибіркові дисперсії є оцінками однієї генеральної сукупності, тобто однорідні.

Якщо $G \geq G_{a(k,n)_т}$, тобто перевірка на відтворюваність дала негативний результат, то на місці грубого промаху виконується додатковий замір, а при неможливості його виконання ставиться середня арифметична повторення рядка.

Процес розрахунків ясно з формул, наведених на полі табл. 4 і додаткового пояснення не потребує.

Так як k=14 і n=3 маємо $G < G_{05(14,\ 2)_т}$, тобто 0,1003 < 0,3517, усі вибіркові дисперсії на 0.05 рівні значущості є оцінками однієї генеральної сукупності, тобто однорідні.

Визначення коефіцієнту ***r*** лінійної кореляції Пірсона у разі отримання лінійних

регресійних рівнянь краще виконувати за допомогою ППП Statistica, який видає більш інформативні результати порівняно з програмою Excel. Але ми аналізуємо вихідні дані рівняння функції відгуку другого порядку, тому необхідно отримати кореляційне відношення **h**, яке характеризує криволінійну кореляцію, тобто тісноту зв'язку результуючою ознаки з фактором і змінюється від 0 до 1.

Найбільш просто це зробити виконавши дисперсійний аналіз вихідних даних за допомогою інструментів Excel чи Statistica. Далі використовуючи таблицю його результатів (чарунки A1:F5 табл. 5) за формулами (4), наведених на її полі, визначається індекс детермінації фактора h^2, а потім кореляційне відношення h.

$$\eta^2_{\text{ф.п}} = \frac{SS_{\text{ф.п}}}{SS_{\text{р}}} \quad \text{і} \qquad \eta_{\text{ф.п}} = \sqrt{\frac{SS_{\text{ф.п}}}{SS_{\text{р}}}}, \tag{4}$$

де $SS_{\text{ф.п}}$ – сума квадратів відповідного фактора чи їх поєднання;

$SS_{\text{Р}}$ – загальна сума квадратів.

Результати розрахунку за формулами (4) подано в табл. 5.

З неї бачимо сильне кореляційне відношення між залежним фактором – щільністю брикетів і довжиною часток та помітне (за шкалою Чеддока) з умістом зв'язуючої речовини. Між щільністю брикетів і конусністю матриці наявне слабке кореляційне відношення. Про істотність кореляційного відношення свідчить не його величина, а рівень значущості [6]. Для всіх названих кореляцій за значеннями поданими в колонці F табл. 5 він є статистично значущим на прийнятому рівні а.

Таблиця 5 – Визначення кореляційного відношення досліджуваних факторів

	A	B	C	D	E	F	G	H
1	Джерело варіювання	Сума квадратів SS	Число ступенів вільності df	Середній квадрат MS	Критерій Фішера F_ϕ	Рівень значущості p	Індекс детермінації η^2	Кореляційне відношення η
2	Вільний член	*5102389*	*1*	*5102389*	*1995347*	*0.000000*		
3	Довжина частки *l*, мм	*16333*	*2*	*8166*	*3194*	*0.000000*	*0.688*	*0.829*
4	Уміст зв'язуючого δ, %	*7022*	*2*	*3511*	*1373*	*0.000000*	*0.296*	*0.544*
5	Конусність матриці α, град	*373*	*2*	*186*	*73*	*0.000021*	*0.016*	*0.125*
6	Похибка	18	7	3	G3=B3/СУММ(B$3:B$5)			H3=G3^0.5

Між незалежними факторами – довжиною часток соломи, умістом зв'язуючої речовини та конусністю матриці брикетного пресу кореляційні відношення відсутні, що забезпечується самою організацією плану активного експерименту.

Висновки. Наведені способи попередньої оцінки і обробки отриманих експериментальних даних при регресійному аналізі, які ґрунтуються на використанні Microsoft Excel і пакету прикладних програм Statistica, дають змогу без залучення довідкових таблиць показників розподілу оціночних критеріїв визначити і оцінити всі необхідні характеристики вибірки, і, відповідно, її придатність для регресійного аналізу.

Незначна модифікація наведених таблиць дає змогу легко отримувати високоточні результати при використанні матриць планування для будь якої кількості факторів.

Список літератури

1. Саутин С. Н. Планироване эксперимента в химии и химической технологии / С. Н. Саутин. – Л. Химия, Ленингр. отд., 1975. – 48 с.
2. Любченко Е. А. Планирование эксперимента: учебное пособие. Часть 1 / Е. А. Любченко, О. А. Чуднова. Владивосток: Изд-во ТГЭУ, 2010. -156 с.
3. Лапач С. М. Статистичні методи в медико–біологічних дослідженнях із застосуванням Excel / С. М. Лапач, А. В. Губенко, П. М. Бабіч – 2–е вод., перероб. і доп. – К.: МОРІОН, 2001. – 408 с.

4. Бакарджиев Р. А. Обоснование конструктивных параметров и режимов работы пресс-брикетировщика для утилизации растительных материалов: дисс…канд. техн. наук: спец. 05.20.01 / Бакарджиев Роман Александрович / Мелитополь, 1997. – 168 с.
5. Веденяпин Г. В. Общая методика экспериментальных исследований и обработки опытных данных / Г. В. Веденяпин – М.: Колос, 1973. – 199 с.
6. Подання результатів математичної та статистичної обробки даних медичних та біологічних досліджень у дисертаційних роботах / [Сердюк А. М., Антомонов М. Ю., Бардов В. Г., Прилуцький О. С.] // Бюлет. Вищої атестац. комісії України. – 2010. – № 6. – С. 31–33.

Roman Bakardzhyev
Tavria State Agrotechnical University
Andrew Komarov
Taras Shevchenko National University of Kyiv
Preliminary assessment and data processing in regression anlysis

Methods of preliminary assessment and data processing during the regression analysis of the active experiment with application of Excel statistical functions and Statistica software package are represented.

Sampling test for the statistical distribution normality , evaluation of equivocal results by the Student's t -test and checking the results reproducibility by Cochran's test are performed without using of the reference evaluation criteria distribution tables, mutual correlation ratio of dependent and independent features are determined and given by specific examples.

These methods allow to assess the sampling with any number of factors by the regression analysis suitability fast and highly accurately.
regression analysis, statistical distribution normality, Student's t -test, Cochran's test, correlation ratio

Одержано 12.03.15

УДК 681.513.5

Б.М. Гончаренко, проф., д-р техн. наук, О.П. Лобок, доц., канд. фіз.-мат. наук
Національний університет харчових технологій, ladanyuk@nutt.edu.ua
Л.Г. Віхрова, проф., канд. техн. наук
Кіровоградський національний технічний університет

Робастне керування нелінійними об'єктами з запізнюванням

В роботі шукається робастне керування у вигляді зворотного зв'язку від стану нелінійної динамічної системи, що функціонує в умовах параметричної невизначеності та наявності запізнювання. Проблема нелінійності об'єкта керування при наявності запізнювання впливає на характер керованих динамічних процесів і суттєво впливає на вигляд та складність шуканих керувань. Тому важливо, що наведена схема формування керувальної дії дозволяє з заданою точністю не тільки відслідковувати заданий еталонний сигнал, але і враховувати ефект запізнювання, наявність нелінійності в умовах апріорної невизначеності та дії зовнішніх збурень.

Пропонується структура робастної системи керування нелінійним динамічним об'єктом з запізненням за станом, яка дозволяє компенсувати параметричну невизначеність і зовнішні обмежені збурення з заданою точністю. Для цього запропоновано алгоритм формування сигналу, за допомогою якого проводиться оцінка збурень і формується керування, що забезпечує необхідну динамічну точність.
робастне керування, нелінійний динамічний об'єкт, запізнювання, параметрична невизначеність, зовнішні збурення, еталонна модель