

УДК 519.237.5: 621.9

А.В. Мигович¹, С.М.Лапач¹

¹ – Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», м. Київ, Україна

Деякі помилки в роботі Microsoft Excel

Табличний процесор Microsoft Excel є масовим і могутнім засобом для підтримки розрахунків практично в будь якій сфері життя [3, 5]. Разом з тим необхідно застерегти від сприйняття отриманих за допомогою нього результатів як істини в останній інстанції. Мова йде не тільки про помилки і недоліки в функціях статистичної обробки в Microsoft Excel, які викликані недостатньою обізнаністю зі статистичними методами тих, хто ці функції створював [1, 2], що приводить до неправильного застосування методів математичної статистики, і не про не врахування нестійкості деяких процедур чисельних методів (розв'язання систем лінійних рівнянь, обернення матриця) [6], результатом чого є отримання невірних результатів як наслідок цієї нестійкості. Мова також йде про програмні помилки, походження яких поясненню стороннім спостерігачем не піддається і які через те є особливо небезпечними.

435

В роботі приводиться приклад задачі по побудові регресійної моделі, яка засобами Microsoft Excel несподівано розв'язується абсолютно неправильно (виявлена випадково при виконанні студентських дослідницьких робіт). При цьому особливо небезпечним є те, що в результаті, який продукує програма присутні як абсолютно правильні, так і абсолютно неправильні розрахунки, що, звісно, може привести до пропуску помилки і прийнятті в роботу неправильних розрахунків.

Вихідні дані представлені в табл.1. Це матриця повного факторного експерименту (ПФЕ) виду $3^1 \times 4^1 // 12$. ПФЕ є ідеальним варіантом з точки зору теорії планування експериментів за усіма показниками: всі статистичні показники, які розраховуються незалежні і незміщені, обчислювальна стійкість ідеальна, структурна стійкість також. Деякі можливі проблеми з протиріччями в перевірці гіпотез в ПФЕ [7] не стосуються теми роботи і не приймаються до

уваги. В даній роботі розглядається тільки обчислення регресійних коефіцієнтів.

Таблиця 1. Міцність деревини в залежності від вологості і температури

№ експ.	Вологість	Температура	Міцність					Середнє
1	15	20	385	333	326	323	344	342,2
2	15	35	354	330	306	293	277	312
3	15	50	255	242	238	222	215	234,4
4	15	100	149	143	127	122	109	130
5	30	20	226	221	205	193	193	207,6
6	30	35	185	182	168	156	151	168,4
7	30	50	148	132	125	112	89	121,2
8	30	100	82	68	68	61	44	64,6
9	100	20	230	206	191	186	182	199
10	100	35	212	182	175	152	147	173,6
11	100	50	144	128	115	114	109	122
12	100	100	91	64	59	54	50	63,6

Застосовуючи програмний засіб ПРИАМ (планування, регресія і аналіз моделі) [4] будуємо регресійну модель.

Отримана модель високоінформативна ($R=0,995695$; $F_R=43,271 > F_{кр}=2,126$; параметр критерію Бокса-Веца $\gamma=4$), адекватна ($F_{ад}=1,655 < F_{кр}=4,066$), з добрим описом експериментальних даних (Середній % відхилення 3,02), обчислювально ($cond=1$) і структурно стійка (матриця регресорів ортогональна). Тобто, ніяких проблем в побудові і використанні цієї моделі не спостерігається.

Вид моделі (без відкидання незначущих коефіцієнтів) представлено нижче.

$$\hat{y} = 178,2167 - 34,6724f_1^1 + 41,46064f_1^2 - 61,0558f_2^1 + 10,21646f_2^2 + 9,228068f_1^1f_2^1 - 0,49237f_1^1f_2^2 - 10,4712f_1^2f_2^1 - 1,73802f_1^2f_2^2,$$

де

$$f_1^1 = 0,019355(X_1 - 48,3333);$$

$$f_1^2 = 4,34571((f_1^1)^2 - 0,44534f_1^1 - 0,51405);$$

$$f_2^1 = 0,020513(X_2 - 52,25);$$

$$f_2^2 = 2,7155669((f_2^1)^2 - 0,45941f_2^1 - 0,38067)$$

В табл.2 представлені статистичні характеристики приведеної регресійної моделі.

Таблиця 2. Коефіцієнти регресії і їх статистичні характеристики

Ім'я	Коефіцієнт	СКв помилка	t-розрахункове	beta-коэф.	Частка впливу
X 1	-34,6724	4,481179	-7,73735	-0,41407	0,171454
X 2	41,46064	4,481179	9,252173	0,495137	0,245161
X 3	-61,0558	4,481179	-13,6249	-0,72915	0,531659
X 4	10,21646	4,481179	2,27986	0,122009	0,014886
X 5	9,228068	4,481179	2,059295	0,110205	0,012145
X 6	-0,49237	4,481179	-0,10988	-0,00588	3,46E-05
X 7	-10,4712	4,481179	-2,3367	-0,12505	0,015638
X 8	-1,73802	4,481179	-0,38785	-0,02076	0,000431
Вільний член	178,2167				

За матрицею ортогональних контрастів побудуємо модель засобами Microsoft Excel. Отриманий результат представлено в табл.3.

Таблиця 3. Результати регресійного аналізу засобами Excel.

<i>Регрессионная статистика</i>	
Множественный R	0,995695
R-квадрат	0,991408
Нормированный R-квадрат	0,968497
Стандартная ошибка	15,52326
Наблюдения	12

Дисперсионный анализ

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>
Регрессия	8	83416,96	10427,12	43,27117
Остаток	3	722,9146	240,9715	
Итого	11	84139,88		

	<i>Коэффициен ты</i>	<i>Стандартная ошибка</i>	<i>t- статистик а</i>	<i>P- Значение</i>
Y-пересечение	178,2167	4,481179	39,77004	3,5E-05
Переменная X 1	0	4,481179	0	1
Переменная X 2	0	4,481179	0	1
Переменная X 3	0	4,481179	0	1
Переменная X 4	0	4,481179	0	1
Переменная X 5	9,228068	4,481179	2,059295	0,131585
Переменная X 6	-0,49237	4,481179	-0,10988	0,919446
Переменная X 7	0	4,481179	0	1
Переменная X 8	-1,73802	4,481179	-0,38785	0,72401

438

Як ми можемо бачити, статистичні характеристики моделі практично повністю співпадають з відповідними характеристиками отриманої раніше моделі. Але з коефіцієнтів регресії співпадають тільки чотири з дев'яти. Всі інші дорівнюють нулю. Що цікаво, дорівнюють нулю зовсім не самі слабкі за впливом на відгук. Щоб у цьому впевнитись достатньо подивитись на значення в стовпчику «Частка впливу» в табл..2.

Звичайно, ми можемо отримати регресійні коефіцієнти і іншим способом, використовуючи не надбудову «Анализ данных», а вбудовані статистичні функції табличного процесора.

Використовуємо функцію =ЛИНЕЙН(\$K3:\$K14;C3:C14) і отримуємо матрицю коефіцієнтів (див табл.. 4).

Таблиця 4. Коефіцієнти регресії, отримані функцією =ЛИНЕЙН(\$K3:\$K14;C3:C14)

-34,6724	0	0	10,21646	9,228068	0	-10,4712	0
----------	---	---	----------	----------	---	----------	---

Як ми бачимо, результати співпадають з роботою надбудови «Анализ данных» і не відповідають тестовим значенням, отриманим з використанням ПРИАМ.

Оскільки в нас матриця ортогональна, то ми можемо виконати обчислення, використовуючи вбудовані формули для роботи з масивами без небезпеки отриманні невірних результатів [6].

Коефіцієнти регресії в такому випадку визначаються за формулою $B = (X^T X)^{-1} X^T Y$. Результат обчислень приведений в табл..5.

Таблиця 5. Результат, отриманий при обчисленні за формулами

B=	-34,67243286
	41,46063943
	-61,05579037
	10,21646174
	9,228068311
	-0,492369528
	-10,47116296
	-1,738023872

От цей результат співпадає з тестовим і не співпадає з отриманим за спеціалізованими статистичними функціями і надбудовою табличного процесора. Тобто, функція «Регрессионный анализ» надбудови «Анализ данных» містить помилку, яку важко пояснити виходячи зі звичайних способів розрахунку коефіцієнтів регресії.

Як можна виявити таку помилку? Звернемо увагу на регресійні залишки (табл..6). При коефіцієнті множинної кореляції 0,995695 і частці загального розсіювання, яка пояснюється рівнянням регресії 96,85% (див. табл..3) вони занадто великі: практично на рівні самих значень модельного відгуку. Дійсно,

якщо ми «вручну» перерахуємо залишкову дисперсію за даними табл. 6, то отримаємо значення 27692,94, а зовсім не 240,97, яке міститься в табл.3. Така невідповідність є сигналом про помилки в отриманих результатах і вимагає роботи по їх виявленню. Наприклад, як показано вище, перерахункам іншим способом чи за допомогою інших програмних засобів.

Тобто, необхідно проаналізувати отримані результати на наявність в них протиріч, невідповідностей між різними показниками і результатами при побудові регресійної моделі.

Таблиця 6. Регресійні залишки за Microsoft Excel

ВЫВОД ОСТАТКА

<i>Наблюдение</i>	<i>Предсказанное</i>	<i>Y</i>	<i>Остатки</i>
1	185,061	157,139	
2	183,3421	128,6579	
3	180,5836	53,81643	
4	163,8799	-33,8799	
5	186,0859	21,51409	
6	179,564	-11,164	
7	174,8638	-53,6638	
8	172,3529	-107,753	
9	163,5031	35,49694	
10	171,7438	1,856157	
11	179,2026	-57,2026	
12	198,4172	-134,817	

Перевірка показала, що в версіях, починаючи з 2010, ця помилка виправлена і модель рахується правильно, тобто результат відповідає тестовим результатам, отриманим на програмному засобі ПРІАМ чи прямими розрахунками функціями для роботи з масивами в Microsoft Excel.

Висновки

1. В статистичних функціях і надбудовах Microsoft Excel існують помилки, в тому числі і такі, що роблять результати непридатними для використання.
2. В деяких випадках ці помилки не можуть бути виявлені простим аналізом правильності використання статистичних методів, бо вони є результатом внутрішніх помилок в програмному забезпеченні.
3. Рекомендація перераховувати іншими засобами рівноцінна відмові від Microsoft Excel, або збільшенні зусиль в кілька разів і втраті зручності, що є неприйнятним.
4. Залишається аналіз результатів (а не бездумне їх прийняття для розміщення у звіті) і виявлення можливих протиріч. При наявності протиріч уже потрібні обчислювальні їх перевірки, і, при підтвердженні, перерахунки іншими засобами. Що, відверто кажучи і є звичайною роботою дослідники при обробці даних.

Список використаних джерел

1. Карлсберг К. Регрессионный анализ в Microsoft Excel –СПб.: Альфа книга, 2017. –400с.
2. С.Н. Лапач, А.В. Чубенко, П.Н. Бабич Статистика в науке и бизнесе –К.: 2002, Морион. – 640с.
3. Берк К., Кейри П. Анализ данных с помощью Microsoft Excel –М.: Вильямс, 2005. –560с.
4. Лапач С.Н., Радченко С.Г., Бабич П.Н. Планирование, регрессия и анализ моделей PRIAM (ПРИАМ) / Каталог программные продукты Украины. К.: 1993. С. 24-27.
5. Минько А.А. Статистический анализ в MS Excel –М.: Вильямс, 2004. –448с.
6. Огороднік С.В., Лапач С.М. Поліноми Чебишева і звичайні поліноми в регресії при використанні масових програмних засобів / Загально університетська науково-технічна конференція молодих вчених та студентів, Київ. 2012р. С. 105–107.
7. С.Н. Лапач, С.Г. Радченко Основные проблемы построения регрессионных моделей // Математичні машини і системи, 2012, № 4, С. 125–133.