УДК 519.237.5: 621.9
Р.Ю. Найчук[1], С.М. Лапач[1]
[1] – Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», м. Київ, Україна

Невиявлена гетероскедастичність і її наслідки

Наявність гетероскедастичності (неоднорідності дисперсій) є порушенням однієї з передумов регресійного аналізу, а саме передумови про постійність дисперсії помилки. Саме цьому порушенню приділено найбільше уваги в науковій літературі, так як воно вважається, с одного боку, найбільш поширеним, з другого, найбільш небезпечним в зв'язку з чутливістю метода найменших квадратів. При її наявності необхідна корекція регресійної моделі [2, 4, 5].

Для активних експериментів з наявністю повторних дослідів для перевірки однорідності дисперсій, як правило, використовують критерій Кохрена [1]. Для визначення наявності гетероскедастичності при відсутності повторних дослідів існує досить багато методів [3, 5]. Як альтернативний варіант пропонується побудова прямолінійної регресійної залежності середньоквадратичної похибки за експериментами від середнього в цих експериментах і прийняття рішення про наявність гетероскедастичності по значимому коефіцієнту при куті нахилу [7].

В реальних дослідженнях ситуація в багатьох випадках складніша, ніж це описано в відповідних підручниках і монографіях, і вимагає від дослідника прийняття рішень, обґрунтування яких лежить за межами просто математичних розрахунків. Розрахунки і перевірка відповідних статистичних критеріїв слугують лише вихідними даними, опираючись на які, дослідник, використовуючи знання з предметної галузі, в якій відбувається дослідження, а також всю іншу доступну йому інформацію, приймає рішення про рівень впливу наявного відхилення від передумов регресійного аналізу і необхідності коригування математичної моделі. Розглянемо ці складності і питання, які при цьому виникають, на прикладі простої задачі моделювання обробки металів різанням.

Вихідні дані приведені в табл..1.

© Р.Ю. Найчук, С.М. Лапач 2021

Таблиця 1. Робоча матриця і результати дослідів

Номер досліду	Незалежні змінні			Результати експериментів			Y середнє	Дисперсія
	X_1	X_2	X_3	Y_{11}	Y_{12}	Y_{13}		
1	56	0,049	0,5	20,87912	30,95238	27,04913	26,29354	25,79582
2	226	0,049	0,5	7,763975	9,037901	7,348243	8,05004	0,775111
3	56	0,2	0,5	27,55102	26,59794	19,14414	24,43103	21,19052
4	226	0,2	0,5	5,30303	5,776173	7,119741	6,066315	0,888246
5	56	0,049	2	32,27092	47,39677	35,50864	38,39211	63,43364
6	226	0,049	2	11,84211	17,93103	10,84011	13,53775	14,72669
7	56	0,2	2	28,83075	40,11194	33,1383	34,027	32,40864
8	226	0,2	2	4,347826	3,846154	4,83871	4,34423	0,246301
9	0	0	0	25,77963	16,11384	4,84211	15,57853	109,8099

Перевірка за критерієм Кохрена показує відсутність гетероскедастичності: $G_{розрахункове}=0,407798<G_{критичне}=0,477494$ при рівні значущості $\alpha=0,05$. Тобто, виходячи з цієї перевірки в нас не має «викидів» і дисперсію помилки можна вважати константою.

Побудована регресійна функція залежності середньоквадратичної похибки від відгуку $y = 1,08428 + 0,76996x$ неінформативна ($F_{розрахункове}=5,057996<F_{критичне}=5,591448$) і з не значимим коефіцієнтом кута нахилу ($t_{розрахункове}=2,248999<t_{критичне}=2,262157$) при рівні значущості $\alpha=0,05$. Звернемо тут увагу на той факт, що в зв'язку з особливостями регресійного аналізу можливі конфлікти і протиріччя між результатами перевірки різних гіпотез [11]. В даному випадку, між інформативністю моделі і значимістю коефіцієнта регресії, що приводить до неоднозначності висновків про гетероскедастичність. У нашому прикладі все виглядає благополучно.

Розглянемо тепер графік залежності, за якою ми будували регресію (рис.1).

© Р.Ю. Найчук, С.М. Лапач 2021

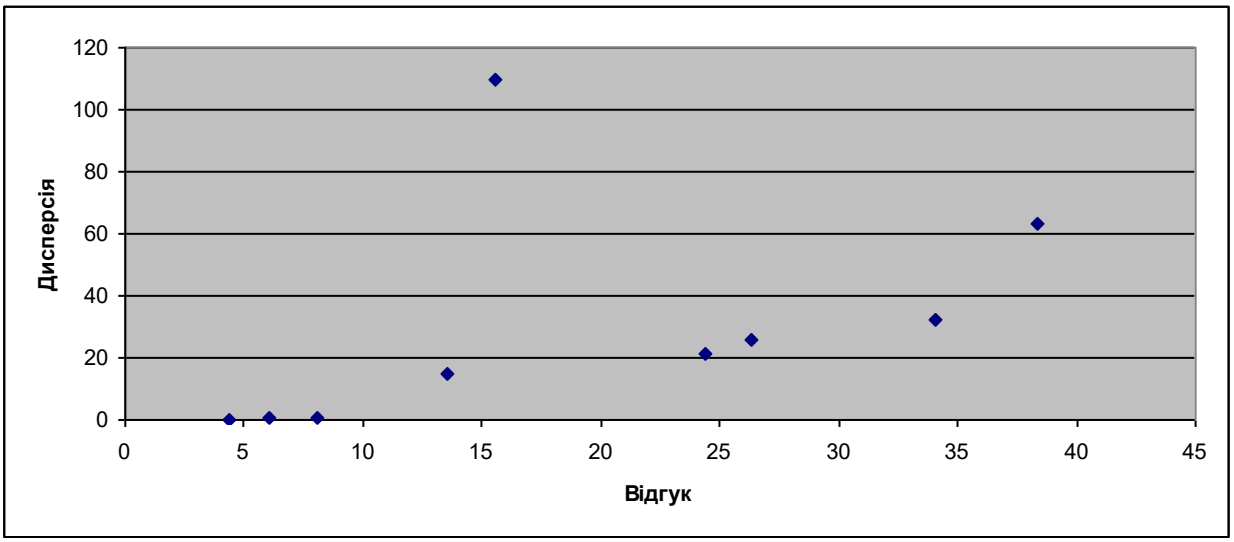

Рис. 1. Залежність дисперсії від відгуку

З рисунку можна зробити висновок про наявність викиду і залежності дисперсії від відгуку, тобто гетероскедастичності. Але ж незалежні перевірки ніби то свідчать про відсутність обох явищ. Проведемо дослідження, в якому порівняємо моделі, побудовані за вказаними даними при наступних припущеннях:

1) порушення передумов відсутні;

2) наявний «викид», який корегується заміною значення, яке найбільш виділяється, на середнє в експерименті;

3) наявна гетероскедастичність, в зв'язку з чим використовується зважений метод найменших квадратів.

Відповідні моделі та їх статистичні характеристики приведені в табл..2, а в табл.. 3 – зміст умовних позначень. Розрахунки моделей виконувались як засобами Microsoft Excel [6, 9] так і ПРІАМ [8]. Як видно, кожне наступне припущення приводить до моделей з вищою інформативністю (величина і значимість коефіцієнта множинної кореляції) і точністю опису даних. Крім того, досить сильно змінюються і значення коефіцієнтів регресії при X_2 і X_3. Останнє легко пояснюється з табл..4. Вказані регресори слабко закорельовані з відгуком і сильно між собою, що приводить до нестійкості обчислювальних процедур і «перетікання» впливу між регресорами. Зміна інформативності і точності опису наскільки велика, що її не можна ігнорувати в прикладних дослідженнях.

Допускаючи, що обидва припущення («викид» і гетероскедастичність)

© Р.Ю. Найчук, С.М. Лапач 2021

підтверджуються, проведемо аналіз вихідних даних, повернувшись до табл..1 і рис.1. З табл..1 видно, що «викидом» є нульова точка, яка знаходиться взагалі за межами основного експерименту.

Таблиця 2. Статистичні характеристики моделей побудованих за всіма експериментами плану

Характеристика	Позначення	Вихідна модель	Модель з корегуванням	Зважена регресія
Інформативність	R	0,886808	0,937188	0,995146
	F_R	6,137	12,031	170,438
	$F_{кр}$	3,028	3,028	3,028
Значимість коефіцієнтів	b_0	25,31694	29,05915	36,85727
	t_0	4,774	7,271	104,073
	b_1	-0,11028	-0,11818	-0,12139
	t_1	-3,993	-5,678	-20,438
	b_2	-2,22873	-11,0661	-18,2418
	t_2	-0,072	-0,472	-2,725
	b_3	6,948475	6,049319	-0,60144
	t_3	2,222	2,567	-0,894
	$t_{кр}$	2,101	2,101	2,101
Адекватність	$S^2_{ад}$	52,45608	29,78791	137,5565
	$S^2_{відт}$	29,91942	21,17861	29,91942
	$F_{ад}$	1,753	1,509	4,598
	$F_{кр}$	9,013	9,013	9,013
Точність	%	47,22	37,83	29,94
Рівень значущості	α	0,05		

Таблиця 3. Умовні позначення в табл..2 і 5.

Умовне позначення	Назва параметру
R	Множинний коефіцієнт кореляції
F_R	Розрахункове значення критерію Фішера для перевірки значимості R

© Р.Ю. Найчук, С.М. Лапач 2021

$F_{кр}$	Критичне значення критерію Фішера для перевірки значимості R
b_0	Вільний член регресійної моделі
t_0	Розрахункове значення критерію Стюдента для перевірки значимості коефіцієнта регресії b_0
b_1	Коефіцієнт при X_1 в регресійній моделі
t_1	Розрахункове значення критерію Стюдента для перевірки значимості коефіцієнта регресії b_1
b_2	Коефіцієнт при X_2 в регресійній моделі
t_2	Розрахункове значення критерію Стюдента для перевірки значимості коефіцієнта регресії b_2
b_3	Коефіцієнт при X_3 в регресійній моделі
t_3	Розрахункове значення критерію Стюдента для перевірки значимості коефіцієнта регресії b_3
$t_{кр}$	Критичне значення критерію Стюдента для перевірки значимості коефіцієнтів регресії
$S^2_{ад}$	Дисперсія адекватності
$S^2_{відт}$	Дисперсія відтворюваності
$F_{ад}$	Розрахункове значення критерію Фішера для перевірки адекватності
$F_{кр}$	Критичне значення критерію Фішера для перевірки адекватності
%	Середня точність опису даних в процентах відхилення
α	Рівень значущості

Таблиця 4. Аналіз мультиколінеарності і сили впливу регресорів

Регресор	Частка впливу	Таблиця мультиколінеарності		
		Коеф. кор. з У	Макс. коеф. кор. з іншим	Ім'я
X_1	0,570096	-0,75507	0,235	X_3
X_2	0,010909	-0,10445	0,233932	X_3
X_3	0,075277	0,274376	0,235	X_1

Видаляємо цю точку з матриці експерименту і будемо працювати з тими, що залишились. Перевірка однорідності за Кохреном підтверджує відсутність

© Р.Ю. Найчук, С.М. Лапач 2021

гетероскедастичності: $G_{розрахункове}=0,39779 < G_{критичне}=0,515687$ при рівні значущості $\alpha=0,05$.

А ось побудована регресійна функція залежності середньоквадратичної похибки від відгуку $y=-0,14943+0,19783x$ є високо інформативною ($F_{розрахункове}=90,46479 > F_{критичне}=5,987378$) і з значимим коефіцієнтом кута нахилу ($t_{розрахункове}=9,511298 > t_{критичне}=2,306$) при рівні значущості $\alpha=0,05$. Це означає наявність статистично значимої залежності між відгуком і дисперсією, тобто порушення передумови дисперсії помилки константи. Таким чином, маємо конфлікт висновків за різними перевірками.

Побудуємо моделі за цими експериментами звичайною та зваженою регресіями. Результати приведені в табл..5 і 6. Звертаємо увагу, що модель зваженої регресії має значно вищі показники інформативності і точності опису, що змушує приймати припущення про наявність гетероскедастичності.

Таблиця 5. Статистичні характеристики моделей з 8 дослідами (без нульової точки)

Характеристика	Позначення	Модель стандартна	Модель зважена
Інформативність	R	0,975925	0,997216
	F_R	26,695	238,481
	$F_{кр}$	3,098	3,098
Значимість коефіцієнтів	b_0	36,5754	39,56884
	t_0	8,948	200,019
	b_1	-0,13404	-0,1327
	t_1	-8,477	-23,707
	b_2	-28,816	-19,2323
	t_2	-1,619	-3,052
	b_3	4,243359	-0,59991
	t_3	2,368	-0,946
	$t_{кр}$	2,120	2,120
Адекватність	$S^2_{ад}$	14,45102	63,40526
	$S^2_{відт}$	19,93312	19,93312
	$F_{ад}$	1,379	3,1809

© Р.Ю. Найчук, С.М. Лапач 2021

	$F_{кр}$	9,117	9,117
Точність	%	26,00	17,22

Таблиця 6. Аналіз мультиколінеарності для моделей за 8 експериментами..

Регресор	Коеф. кор. з У	Макс. коеф. кор. з іншим	Ім'я
X 1	-0,92444	0	з усіма
X 2	-0,17653	0	з усіма
X 3	0,258229	0	з усіма

Висновки

1. Як уже відмічалось [10] застосування в задачах моделювання процесів різання стандартних процедур регресійного аналізу не завжди можливе, що викликано високим рівнем випадкових факторів і неоднорідністю факторного простору [12].

2. Перевірка за критерієм Кохрена не завжди дозволяє визначити наявність грубих помилок («викидів») і гетероскедастичності.

3. Побудова і аналіз регресійної залежності між середньоквадратичною похибкою і відгуком більш чутливе до наявності гетероскедастичності, ніж критерій Кохрена, хоч і не є панацеєю.

4. Невиявлена гетерескедастичність приводить в результаті до побудови моделей, гірших з точки зору їх практичного застосування в зв’язку з відсутності корекції за допомогою зваженого методу найменших квадратів.

5. Для гарантованою побудови регресійних моделей з задовільними властивостями з точки зору практичного використання при підозрі на наявність гетероскедастичності недостатньо формальних процедур перевірки. Необхідний смисловий аналіз задачі з висуванням припущень про наявність «викидів» і гетероскедастичності і їх перевіркою побудовою відповідних регресійних моделей.

Список використаних джерел

1. Большев Е.Н., Смирнов Н.В. Таблицы математической статистики. Изд. 3-е – М. Наука. ГРФМЛ., 1983. –416с.

2. Вучков И., Бояджиева Л., Солаков Е. Прикладной регрессионный анализ –М.: Финансы и статистика, 1987. 239с.

© Р.Ю. Найчук, С.М. Лапач 2021

3. Доугерти К. Введение в эконометрику –М.: ИНФРА-М, 2001. –402с.
4. Дрейпер, Н. Прикладной регрессионный анализ / Н. Дрейпер, Г. Смит –Изд. 3-е. –М.: Диалектика, 2007. –912с.
5. Грін Г. В. Економетричний аналіз / Г. В. Грін –К.: Основи, 2005. –1198с.
6. Карлберг К. Регрессионный анализ в Microsoft Excel –Спб. ООО»Альфа-книга», 2017. –400с.
7. Кузьмін В.М., Лапач С.М. «Полігональна регресія при наявності гетероскедастичності», «Економіка і управління» –2007, –№1. –С.81–86.
8. 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.
9. С.Н. Лапач, А.В. Чубенко, П.Н. Бабич Статистические методы в медико-биологических исследованиях с использованием Excel –2 изд. перераб. и доп. – К.: 2001, Морион. – 408с.
10. С.М. Лапач Проблеми побудови регресійних моделей процесів різання металів / Вісник НТУУ «КПІ». Серія «Машинобудування» . 2014, №3(72). С.40–47.
11. С.Н. Лапач, С.Г. Радченко С.Г. Основные проблемы построения регрессионных моделей // Математичні машини і системи, 2012, № 4, С. 125–133.
12. А. В. Мигович, С. М. Лапач Неоднорідність факторного простору при моделюванні процесів різання // Збірка праць «Інновації молоді в машинобудуванні» №1 2019 С. 8–15

© Р.Ю. Найчук, С.М. Лапач 2021