Mathematics in Modern Technical University
ISSN 2664-4258

Mathematics in Modern Technical University, 2019(1), 15–23
10.20535/mmtu-2017.7-777

Стійкість коефіцієнта кореляції до «викидів» при використанні в регресійному аналізі

С. М. Лапач

Кафедра технології машинобудування,
КПІ ім. Ігоря Сікорського, Київ, Україна

Анотація

Розглянуто питання стійкості коефіцієнта кореляції за наявності «викидів», які в регресійному аналізі часто є наслідком закону розподілу похибки, відмінного від нормального, наприклад, логнормального чи нормального з «важкими» хвостами. Тоді їх не можна відкинути чи скоригувати і вони залишаються в навчальній вибірці. При цьому відбувається зміщення регресійної моделі в бік відхилень. Крім того, в зв'язку зі зміною коефіцієнтів кореляції внаслідок викидів можлива зміна сформованої структури моделі. Метою роботи є визначення наскільки великим може бути зміщення коефіцієнта кореляції залежно від значення коефіцієнта, способу його обчислення, величини викиду та розміру вибірки для різних коефіцієнтів кореляції.

Ключові слова: кореляційний аналіз; регресійний аналіз; коефіцієнт кореляції; медіанна кореляція.

MSC2010 62-XX

УДК 519.237.5

lapach@ukr.net

Cite as: Lapach, S. M. (2019). Stability of correlation coefficient to "outliers" used in regression analysis. *Mathematics in Modern Technical University, 2019*(1), 15–23. 10.20535/mmtu-2017.7-777

1 Вступ

Коефіцієнти парної кореляції застосовують в регресійному аналізі для визначення конкретної специфікації (також використовують терміни-аналоги: «структура», «оптимальна множина регресорів») (Aivazyan, Buchshtaber, & Yenyukov, 1985; Ezekiel & Fox, 1959; S. M. Lapach, 2017). Особливістю застосування коефіцієнтів кореляції в регресійному аналізі є широке використання малих за абсолютною величиною значень коефіцієнтів, що пов'язано з можливістю множини значущих в сукупності, а не по одному регресорів (Pardoux, 1982). Перевагою парного коефіцієнта кореляції при використанні для визначення конкретної структури рівняння моделі є то, що для «істинних» ефектів він більше, ніж для закорельованих з ним «хибних» (S. N. Lapach, Pasechnik, & Chubenko, 1999). Проблеми, які виникають під час прийняття рішень у регресійному аналізі у зв'язку зі статистичними коливаннями значення коефіцієнта в різних вибірках однієї і тієї ж генеральної сукупності розглянуті в (S. M. Lapach, 2018). В літературі (Orlov, 2018; Shishlyannikova, 2009) відзначається нестійкість цих коефіцієнтів до викидів.

У регресійному аналізі у випадку «викидів», які часто можуть бути просто відхиленням від нормального розподілу помилки, наприклад з «важкими хвостами», які поширені, зокрема, під час випробуваннях міцності, важливо знати, наскільки в цій ситуації можлива деформація структури моделі. Для цього треба знати, наскільки змінюється коефіцієнт кореляції при наявності «викидів».

2 Основні результати

У статті досліджується, наскільки змінюється порівняно з точним випадком оцінка коефіцієнта кореляції при наявності викидів. Аналізується вплив відносного розміру викидів, розміру вибірки, виду коефіцієнта кореляції (Пірсона, Спірмена, Кендала, Пірсона з використанням медіани замість середнього (Mosteller & Tukey, 1977)). За еталон взято значення коефіцієнта кореляції Пірсона для вибірки без викидів. Для дослідження взято штучно створену регресійну модель з різними рівнями залежності відгуку від факторів. Досліджувався вплив одного викиду різного рівня. Еталонні (без похибок і «викидів») значення коефіцієнтів кореляції подано в таблиці 1. Звертаємо увагу на відомий теоретично, але не враховуваний «прикладниками» той факт, що для малих значень коефіцієнтів кореляції їх значення відносно сильно відрізняються при зміні розміру вибірки (див. також в (S. M. Lapach, 2018).

Умовні позначення варіантів вибірок і способів розрахунку коефіцієнтів кореляції подано в таблиці 2.

Тривіальні висновки: стійкість коефіцієнта кореляції, розрахованого будь-яким способом, збільшується при збільшенні розміру вибірки; стійкість зменшується при збільшенні величини викиду.

Табл. 1: Рівень еталонної кореляції

Назва фактору								
Розмір вибірки	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_9
16	0,89588	0,20617	-0,00984	-0,07438	-0,0333	0,02441	0,34488	-0,1525
32	0,85829	0,32919	0,07012	-0,08177	-0,12526	-0,10966	-0,16677	-0,07895

Разом з тим, звертаємо увагу, що переваг використання медіани замість середнього для обчислення коефіцієнта кореляції Пірсона не виявлено, хоча уявлення про таку перевагу досить поширене.

Табл. 2: Умовні позначення в таблицях і на рисунках

Позначення	Розшифрування
П	Коефіцієнт кореляції Пірсона
С	Коефіцієнт кореляції Спірмена
К	Коефіцієнт кореляції Кендала
М	Коефіцієнт кореляції Пірсона з медіаною замість середнього
П+s	Коефіцієнт кореляції Пірсона з викидом розміром σ
С+s	Коефіцієнт кореляції Спірмена з викидом розміром σ
К+s	Коефіцієнт кореляції Кендала з викидом розміром σ
М+s	Коефіцієнт Пірсона з медіаною замість середнього з викидом розміром σ
П+2s	Коефіцієнт кореляції Пірсона з викидом розміром 2σ
С+2s	Коефіцієнт кореляції Спірмена з викидом розміром 2σ
К+2s	Коефіцієнт кореляції Кендала з викидом розміром 2σ
М+2s	Коефіцієнт Пірсона з медіаною замість середнього з викидом розміром 2σ
П+3s	Коефіцієнт кореляції Пірсона з викидом розміром 3σ
С+3s	Коефіцієнт кореляції Спірмена з викидом розміром 3σ
К+3s	Коефіцієнт кореляції Кендала з викидом розміром 3σ
М+3s	Коефіцієнт Пірсона з медіаною замість середнього з викидом розміром 3σ

Таблиця 3 показує залежність тісноти зв'язку сукупності розрахованих коефіцієнтів кореляції для моделі в залежності від розміру вибірки, наявності і величини «викиду», виду коефіцієнта кореляції і наявності випадкової похибки в результатах.

При всіх рівнях «викиду» найближче до еталону відтворюють рівень зв'язку коефіцієнт кореляції Пірсона та його аналог з використанням медіани замість середнього. Відповідність між коефіцієнтами різних типів, як точними, так і з похибками різних рівнів при різних еталонних значеннях коефіцієнта представлено на рис. 1 і рис. 2. Найбільш відрізняється від інших при відсутності викидів коефіцієнт кореляції Кендала, що є досить відомим фактом.

При наявності викидів рівня 2σ (див. рис. 2) відмінності коефіцієнта кореляції Кендала вже не видаються такими великими.

Добре видно, що співвідношення між значеннями коефіцієнтів різних типів залежить від наявності «викидів». Тобто, наявність і величина «викидів» вливає на співвідношення коефіцієнтів кореляції, змінюючи їх. Прийняття рішення про

Табл. 3: Тіснота зв'язку коефіцієнтів кореляції з еталонним

Розмір вибірки N= 16				Розмір вибірки N = 32			
Точна		З похибкою		Точна		З похибкою	
Імена	Рівень кореляції	Імена	Рівень кореляції	Імена	Рівень кореляції	Імена	Рівень кореляції
П	1	П	1	П	1	П	1
М	0,99917	М	0,99971	М	0,99999	М	0,99996
К	0,98597	П+s	0,9839	М+s	0,99511	К	0,99733
С	0,98275	М+s	0,98105	П+s	0,99505	М+s	0,99536
П+s	0,97979	С	0,97508	С	0,99331	П+s	0,9953
К+s	0,97368	С+s	0,96070	С+s	0,99021	С	0,99480
К+2s	0,97368	П+2s	0,94871	К	0,98856	К+s	0,99160
К+3s	0,97368	К	0,94789	С+2s	0,98845	С+s	0,99044
М+s	0,97275	М+2s	0,9433	С+3s	0,98845	К+2s	0,98448
С+s	0,96401	С+2s	0,94044	К+s	0,98265	С+2s	0,98411
С+2s	0,96401	К+s	0,93534	М+2s	0,98138	М+2s	0,98238
С+3s	0,96401	С+3s	0,92369	П+2s	0,98116	П+2s	0,98212
П+2s	0,93354	П+3s	0,90792	К+2s	0,98099	К+3s	0,97609
М+2s	0,92000	К+2s	0,90644	К+3s	0,98099	С+3s	0,97582
П+3s	0,87856	М+3s	0,90040	М+3s	0,96055	М+3s	0,9627
М+3s	0,85975	К+3s	0,89062	П+3s	0,96009	П+3s	0,96215

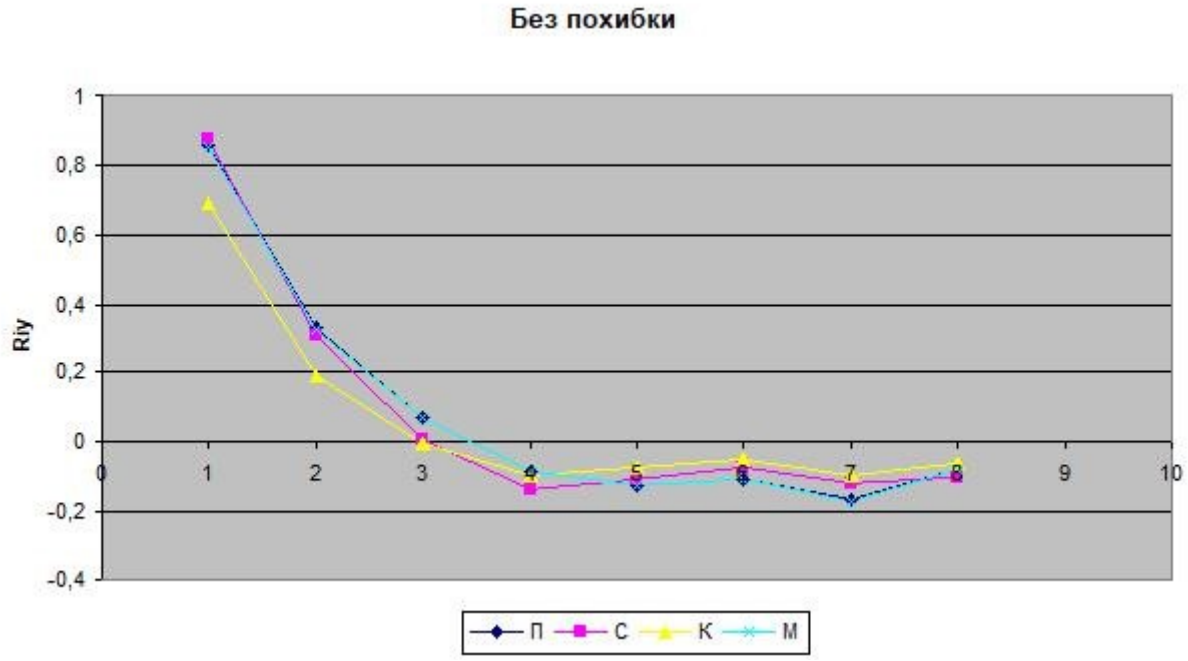

Рис. 1: Відповідність коефіцієнтів кореляції різних видів при відсутності «викидів»

використання певного типу коефіцієнтів кореляції залежить значною мірою від того, наскільки значення кореляції при наявності «викидів» відрізняється від еталонного. Від цього залежить, наскільки спотворюється картина співвідношень рівня зв'язку з відгуком для різних регресорів і структура рівняння регресії. На

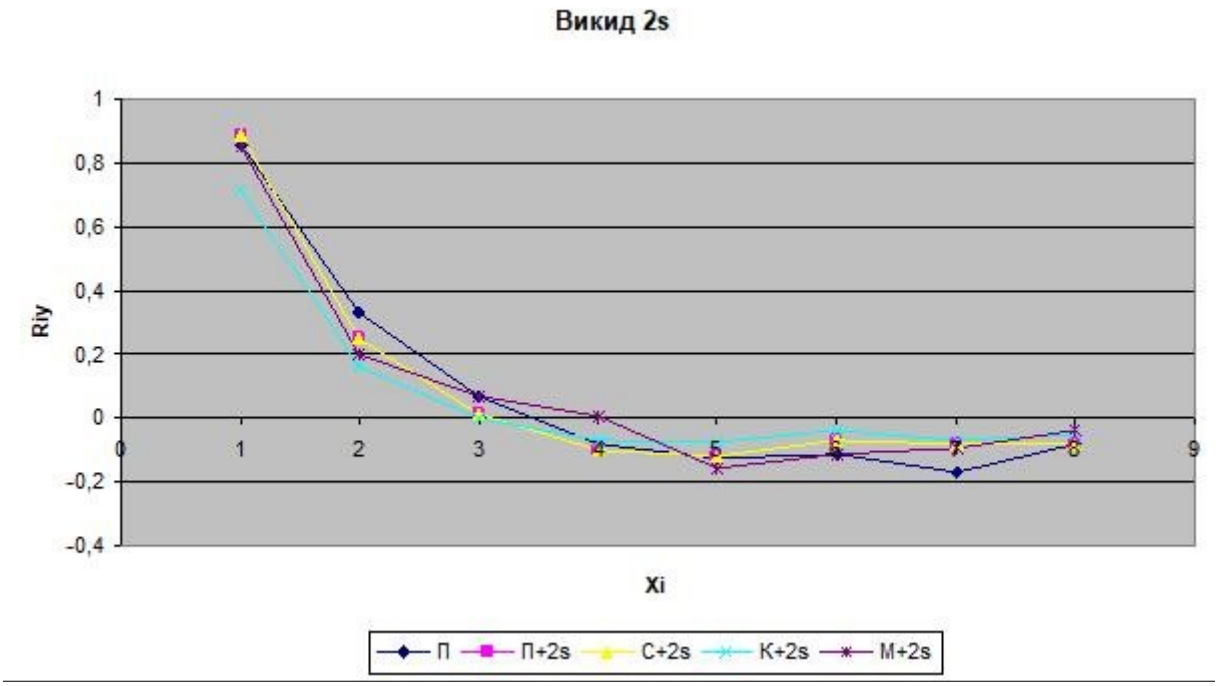

Рис. 2: Відповідність коефіцієнтів кореляції різних видів при наявності «викиду» в 2σ

рис. 3 показано відносний рівень відхилень від еталонного значення для різних коефіцієнтів кореляції при високій закорельованості з відгуком.

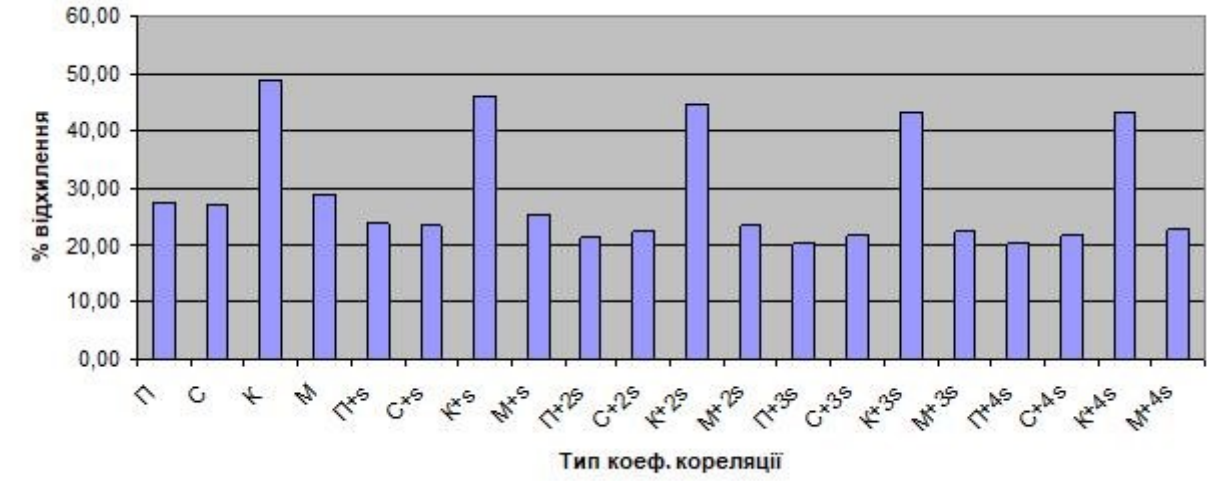

Рис. 3: Відхилення в процентах з похибкою і без для високої кореляції і різних рівнів «викидів»

Найменш стійким виявляється коефіцієнт кореляції Кендала, а стійкість решти досліджуваних коефіцієнтів майже однакова.

Досить несподіваним здається той факт, що відносне відхилення від еталону коефіцієнтів кореляції, розрахованих у точних умовах і з викидами між собою

найбільше при відсутності викидів. Пояснюється це тим, що при наявності викидів уже відбулася достатньо сильна деформація коефіцієнта кореляції.

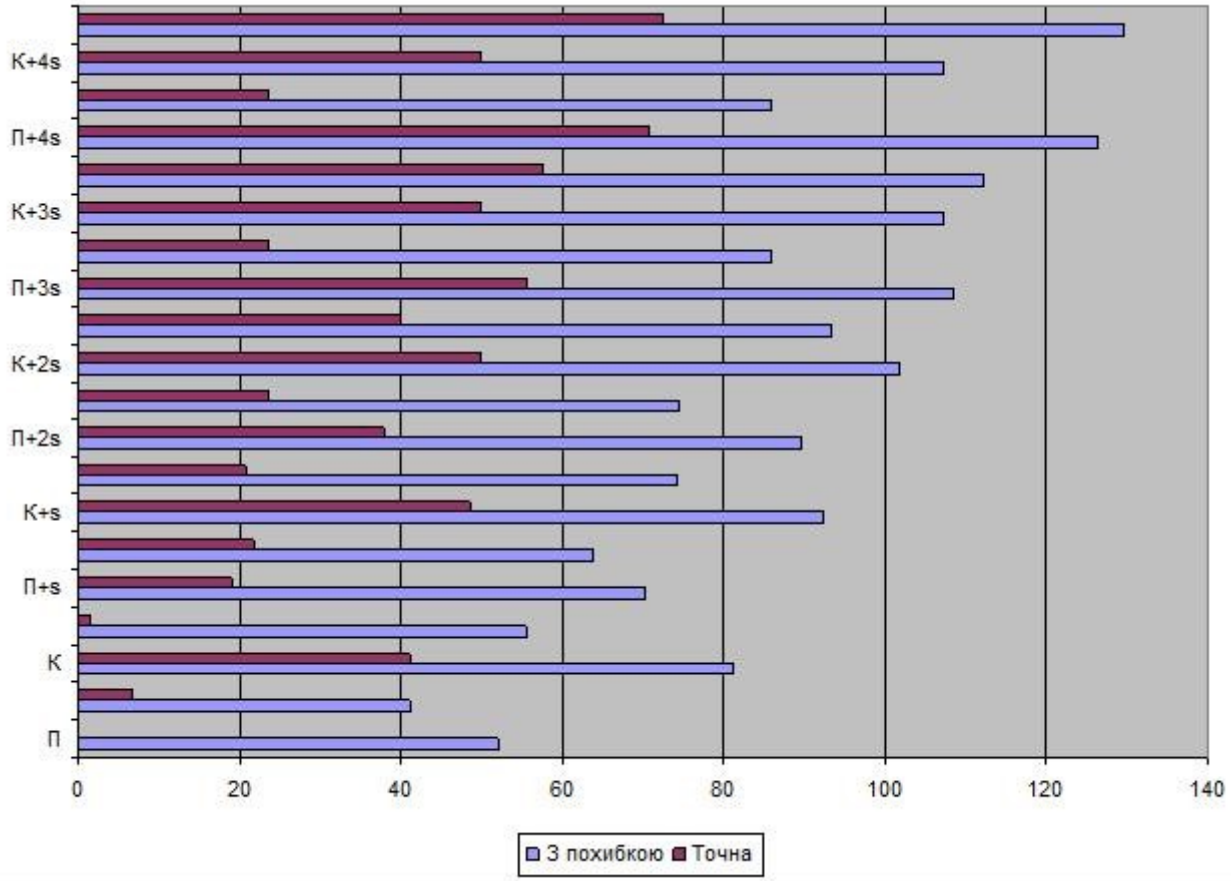

Рис. 4: Відносна похибка для визначення коефіцієнтів кореляції порівняно з еталоном при різних умовах

На рис. 5 показано відносний рівень відхилень від еталонного значення для різних коефіцієнтів кореляції при середній закорельованості з відгуком. Порівнюючи його з ситуацією при високому рівні кореляції (див. рис. 3), бачимо, що, по-перше, рівень відхилень суттєво (майже вдвічі) вище. По-друге, рівень відхилення, на відміну від ситуації з високою кореляцією, зростає відповідно до величини викиду. По-третє, значення відхилення може перевищувати значення самого оригіналу.

Ще однією важливою (з точки зору використання в регресійному аналізі) є структура розподілу коефіцієнтів кореляції з відгуком за своєю абсолютною величиною. Від цієї структури в багатьох алгоритмах формування найкращої підмножини регресорів залежить послідовність вибору претендентів на включення в модель. При наявності мультиколінеарності порушення послідовності приводить до зміщення коефіцієнтів регресії і можливому формування неправильної структури рівняння регресії. У таблиці 4 подано структуру розподілу коефіцієнтів кореляції (ранжовані від більшого до меншого за абсолютною величиною) при різному розмірі викиду і різних коефіцієнтах кореляції. Добре видно, що на своєму місці залишається тільки

найбільший коефіцієнт кореляції. Для інших їх положення в ранжованому ряду змінюються. Чим більше розмір викиду, тим більше може відрізнятись отриманий ряд від оригіналу.

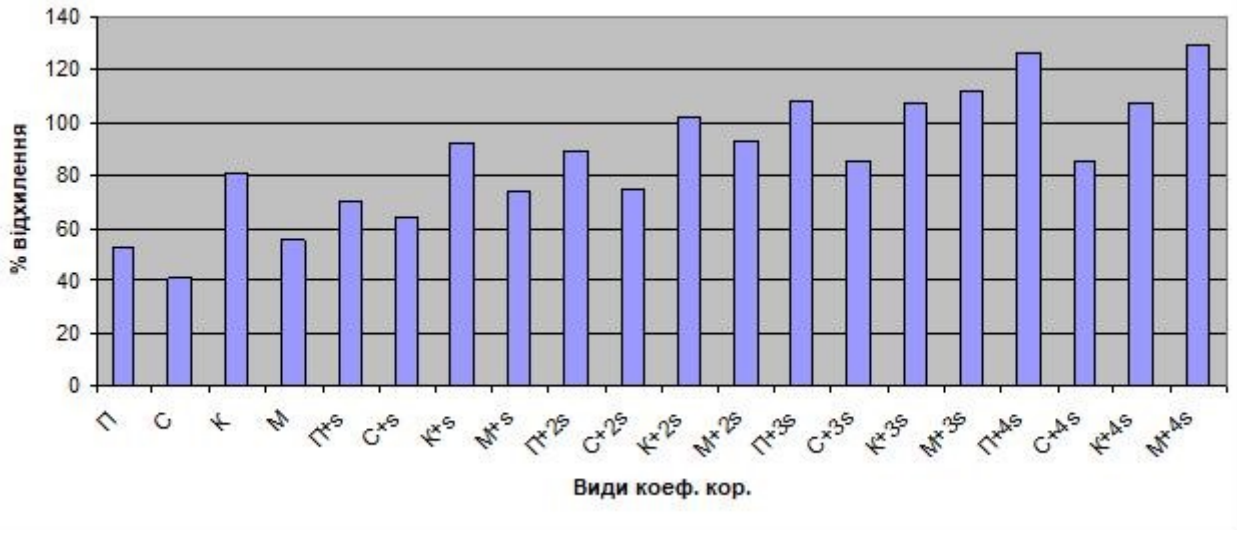

Рис. 5: Відносна похибка для визначення коефіцієнтів кореляції порівняно з еталоном при різних умовах

Табл. 4: Структура розподілу коефіцієнтів кореляції за абсолютною величиною

Вид коефіцієнта кореляції	Назва регресора							
П	X_1	X_2	X_7	X_9	X_6	X_5	X_3	X_4
П+s	X_1	X_2	X_9	X_7	X_6	X_5	X_3	X_4
П+2s	X_1	X_6	X_2	X_5	X_9	X_7	X_4	X_3
П+3s	X_1	X_6	X_5	X_9	X_2	X_7	X_4	X_3
П+4s	X_1	X_5	X_6	X_4	X_9	X_2	X_7	X_3

3 Висновки

1. За стійкістю до «викидів» коефіцієнт кореляції Пірсона з використанням медіани замість середнього практично не відрізняється від традиційного.

2. Найбільш стійкими до викидів є коефіцієнти кореляції Пірсона при всіх рівнях «викидів».

3. Співвідношення рівня відхилень від еталонного значення кореляції різних видів змінюється залежно від рівня викидів.

4. Найбільш відрізняється від еталонного в будь-яких умовах коефіцієнт кореляції Кендала, але при наявності викидів відмінність зменшується.

5. Структура співвідношень між абсолютними значеннями коефіцієнтів не зберігається, за винятком найсильнішого, який таким і залишається. Імовірність

порушення зростає зі зменшеннями абсолютного значення коефіцієнта кореляції і зі зменшенням розміру вибірки.

Отже, гіпотеза про більшу стійкість до «викидів» рангових коефіцієнтів кореляції чи використання медіани замість середнього при обчисленні кореляції не підтверджується. У регресійному аналізі наявність «викидів» при слабкому зв'язку регресорів з відгуком означає збільшення імовірності неправильного визначення структури моделі. Це означає, що при наявності викидів, які не можна видаляти з вибірки, або при неможливості перевірити наявність викидів при можливій наявності їх в вибірці необхідно враховувати можливість неправильного визначення структури моделі. Це означає, по-перше, аналіз структури за допомогою знань, які знаходяться за межами проведеного експерименту: інформація про характер залежності відгуку від конкретних факторів, наявність взаємодій, графічний аналіз поведінки процесу за побудованою моделлю тощо. У практиці автора були випадки, коли регресійна модель коригувалась видаленням регресорів високого порядку і взаємодій, яких на думку спеціалістів предметної галузі не мало бути. Разом з тим, при врахуванні цієї інформації слід бути обережним. Знову ж у практиці автора неодноразово були випадки, коли структура моделі не відповідала уявленням замовника, але виявилась правильною. Такі ситуації часто викликані тим, що багатофакторний експеримент за планом часто проводиться в цілому в умовах, про які інформації в дослідника немає. Тому, по-друге, в відповідальних випадках обов'язкова перевірка моделі на контрольній послідовності дослідів. Ці досліди мають бути проведені в умовах, які дозволяють прийняти обґрунтоване рішення про придатність моделі.

References

Aivazyan, S., Buchshtaber, V. M., & Yenyukov, I. S. (1985). *Applied statistics: Study of relationships [in Russian].* Moscow: Finansy i statistika.

Ezekiel, M., & Fox, K. (1959). *Methods of correlation and regression analysis: Linear and curvilinear* (3rd ed.). John Wiley & Sons, Inc.

Lapach, S. M. (2017). Correlation analysis in application to the definition of the structure of the regression equation [in Ukrainian]. In *Proceedings of the Eighteenth International Scientific Conference Mykhailo Kravchuk conference, Kyiv–Lutsk, October 7–10* (pp. 119–123). Kyiv: Igor Sikorsky Kyiv Polytechnic Institute.
http://matan.kpi.ua/public/files/2017/kravchuk-conf2017/Kravchuk2017-vol2.pdf#page=119

Lapach, S. M. (2018). Risks of using the correlation coefficient for a specific regression model specification [in Ukrainian]. *Mathematical Machines and Systems*, *2018*(3), 142–148.

http://www.immsp.kiev.ua/publications/articles/2018/2018_3/03_2018_Lapach.pdf

Lapach, S. N., Pasechnik, M. F., & Chubenko, A. V. (1999). *Statistical methods in pharmacology and marketing of the pharmaceutical market [in Russian]*. Kyiv: CJSC "Ukrspetsmontazh".

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression: A second course in statistics*. Reading, Mass.: Addison-Wesley.

Orlov, A. I. (2018). Errors in the use of correlation and determination coefficients [in Russian]. *Industrial laboratory. Diagnostics of materials*, *84*(3), 68–72. https://doi.org/10.26896/1028-6861-2018-84-3-68-72

Pardoux, C. (1982). Sur la sélection de variables en régression multiple: une mise au point. *Cahiers du Bureau universitaire de recherche opérationnelle Série Recherche*, *39*, 101–133. http://www.numdam.org/item/BURO_1982__39-40__101_0

Shishlyannikova, L. M. (2009). The use of correlation analysis in psychology [in Russian]. *Psychological Science and Education*, *2009*(1), 98–107. http://psyjournals.ru/psyedu/2009/n1/Shishlyannikova.shtml

---

С. М. Лапач (2019). Стійкість коефіцієнта кореляції до «викидів» при використанні в регресійному аналізі. *Mathematics in Modern Technical University, 2019*(1), 15–23.

Submitted: 2019-04-13
Accepted: 2019-05-12

S. M. Lapach (2019). Stability of correlation coefficient to "outliers" used in regression analysis. *Mathematics in Modern Technical University, 2019*(1), 15–23.

Abstract. The question of the stability of the correlation coefficient in the presence of "emissions", which in the regression analysis is often a consequence of the law of distribution of error, is excellent from the normal, for example, lognormal or normal with "severe cases", is considered. In this case, they can not be rejected or corrected and remain in the training sample. At the same time there is a bias of the regression model in the direction of deviations. In addition, due to the change in the correlation coefficients in the emission factors, a change in the structure of the model is possible. The purpose of the work is to determine how large the displacement of the correlation coefficient can be, depending on the size of the coefficient itself, the method of its calculation, the magnitude of the emission, and the size of the sample for the various correlation coefficients.

Keywords: correlation analysis; regression analysis; correlation coefficient; median correlation.