УДК 004.412

МЕХАНІЗМИ ЗМЕНШЕННЯ МУЛЬТИКОЛІНЕАРНОСТІ МІЖ НЕЗАЛЕЖНИМИ ФАКТОРАМИ ПРИ ПОБУДОВІ НЕЛІНІЙНОЇ РЕГРЕСІЙНОЇ МОДЕЛІ ДЛЯ ОЦІНЮВАННЯ JAVA-ЗАСТОСУНКІВ

Орєхов О.С.[1]; Фаріонова Т.А., к.т.н., доц.[2]
[1,2] Національний університет кораблебудування імені адмірала Макарова
[1,2]Україна, Миколаїв
[1]ORCID 0000-0002-0001-0140; [2]ORCID 0000-0003-3384-4712

Анотація. *Метою роботи є зменшення мультиколінеарності між незалежними факторами за рахунок застосування середніх значень при попередній обробці багатовимірних даних на прикладі побудови п'ятифакторної нелінійної регресійної моделі для обробки інформації з метрик коду JAVA-застосунків. Показано, що поступове видалення викидів із набору емпіричних даних збільшує значення VIF та погіршує стабільність моделі. Застосування середніх значень метрик VMQ, TFQ та CBO відносно суми кількості класів та інтерфейсів дозволило знизити кореляцію між факторами та уникнути критичних значень оцінок VIFs.*
Ключові слова: *мультиколінеарність; VIF; нелінійна регресійна модель; негаусівські дані; JAVA-застосунок.*

Вступна частина. При побудові багатофакторних нелінійних регресійних моделей виникає проблема мультиколінеарності між незалежними факторами в навчальній вибірці, якщо незалежні фактори мають високий рівень кореляції між собою. Високий рівень кореляції між незалежними факторами може привести до збільшення чутливості моделі до випадкових змін у даних, знизити стабільність моделі, призвести до спотворення оцінок ваг параметрів моделі та до зростання стандартних похибок [1].

Для багатофакторних нелінійних регресійних моделей, які будуються із застосуванням методів математичної статистики, нелінійного регресійного аналізу, нормалізації негаусівських даних, та пропонують застосування їх ітераційної обробки [2], існує підвищений ризик наявності або появи високого рівня мультиколінеарністю між незалежними факторами в процесі очищення даних від викидів. Крім того, високий рівень мультиколінеарності при певних умовах є відкладеним і неявним на перших етапах обробки негаусівських багатовимірних даних при застосуванні вище наведених методів.

Метою дослідження є зменшення мультиколінеарності між незалежними факторами за рахунок застосування середніх значень при попередній обробці багатовимірних даних на прикладі побудови п'ятифакторної нелінійної регресійної моделі для обробки інформації з метрик коду JAVA-застосунків.

Основна частина. Авторами із платформи GitHub зібрано навчальну вибірку інформації з метрик коду 286 програмних JAVA-застосунків з відкритим кодом для побудови п'ятифакторної нелінійної регресійної моделі. Для оцінювання рівня мультиколінеарності із навчальної вибірки обрано п'ять метрик як незалежні фактори, та які будуть використовуватись при побудові математичної моделі, а саме: загальну кількість класів (CLS), загальну кількість інтерфейсів (INFC), загальну кількість видимих методів (VMQ), загальна кількість всіх полів класів (TFQ) та загальну оцінка зв'язності між об'єктами (CBO)[3].

Оцінювання рівня мультиколінеарності між незалежними метриками проведено за допомогою коефіцієнту впливу дисперсії (VIFs). Для багатовимірних даних з k незалежними

факторами X_i, i=1,2,...,k, коефіцієнти VIFs представлені діагональними елементами оберненої коваріаційної $k \times k$ матриці. Якщо значення коефіцієнту VIF перевищує 10, то це свідчить про наявність проблем із мультиколінеарністю [4].

Для побудови п’ятифакторної нелінійної регресійної моделі, проведено попередню нормалізацію даних за допомогою багатовимірного перетворення Бокса-Кокса, аналіз незалежних факторів на наявність мультиколінеарності, та ітеративне вилучення одного викиду за одну ітерацію обробки даних, відповідно до методики, запропонованої в роботі [2]. Очищення даних від викидів було проведено за 32 ітерації. В Таблиці 1 наведено значення коефіцієнтів VIFs для вибіркових ітерацій.

Таблиця 1 — Значення коефіцієнтів VIFs для багатовимірних даних CLS, INFC, VMQ, TFQ та CBO при ітеративному очищенні даних від викидів

Ітерація	**Розмір вибірки**	**VIFs для факторів**				
		CLS	**INFC**	**VMQ**	**TFQ**	**CBO**
1	286	16,4	2,3	13,2	10,8	20,9
5	281	16,9	3,5	14,0	10,8	22,4
10	276	17,5	3,6	14,6	12,1	29,0
15	271	18,6	4,6	16,0	13,2	32,0
20	266	18,7	4,6	16,1	13,2	32,3
25	261	19,6	4,6	16,2	13,3	32,2
32	254	25,6	4,5	18,5	15,6	36,4

Таблиця 1 демонструє, що вже початкова навчальна вибірка має суттєві проблеми з мультиколінеарністю для обраних факторів. Крім того, поетапне видалення викидів показало, що оцінки коефіцієнтів VIFs мають тенденцію до зростання на кожній наступній ітерації у порівнянні з попередньою. Остання ітерація, вже для очищеного набору даних від викидів, показує значне підвищення оцінок коефіцієнтів VIFs у порівнянні з першою ітерацією, що дозволяє зробити заключення: поступове очищення вибірки від викидів призводить до збільшення мультиколінеарності між незалежними факторами.

Для вирішення проблеми мультиколінеарності, абсолютні значення метрик VMQ, TFQ та CBO замінені на їх середні значення відносно суми загальної кількості класів та інтерфейсів з програмного проєкту - aVMQ, aTFQ та aCBO відповідно. Для модифікованого набору даних з метриками CLS, INFC, aVMQ, aTFQ та aCBO проведено аналогічну попередню нормалізацію, обробку даних та оцінювання коефіцієнтів VIFs для кожної ітерації. Отримані значення оцінок коефіцієнтів VIFs наведено в Таблиці 2.

Таблиця 2 — Значення коефіцієнтів VIFs для багатовимірних даних CLS, INFC, aVMQ, aTFQ та aCBO при ітеративному очищенні даних від викидів

Ітерація	**Розмір вибірки**	**VIFs для факторів**				
		CLS	**INFC**	**aVMQ**	**aTFQ**	**aCBO**
1	286	2,1	2,3	1,1	1,1	1,6
5	281	2,3	2,3	1,1	1,2	1,1
10	276	3,1	3,2	1,2	1,2	1,2
15	271	3,2	3,2	1,2	1,2	1,2
20	266	3,9	3,9	1,2	1,2	1,3
26	260	4,0	4,0	1,2	1,2	1,3

Отримані оцінки VIFs показали, що використання середніх значень метрик VMQ, TFQ та CBO відносно суми класів та інтерфейсів, дозволяє суттєво знизити рівень мультиколінеарності між незалежними факторами нелінійної регресійної моделі.

Висновки. Проведене дослідження дозволило виявити значний рівень мультиколінеарності між незалежними факторами, які було обрано для побудови п’ятифакторної нелінійної регресійної моделі на базі вибірки інформації з метрик коду. Перехід від метрик VMQ, TFQ та CBO JAVA-застосунків до їх середніх значень відносно загальної кількості класів та інтерфейсів дозволило зменшити рівень мультиколінеарності між незалежними факторами при застосуванні методів математичної статистики, нелінійного регресійного аналізу та ітеративної обробки багатовимірної вибірки інформації з метрик коду JAVA-застосунків.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* New York: Wiley.

[2] Prykhodko, S., & Prykhodko, N. (2021). *Mathematical Modeling of Non-Gaussian Dependent Random Variables by Nonlinear Regression Models Based on the Multivariate Normalizing Transformations*. Mathematical Modeling and Simulation of Systems (MODS'2020), Advances in Intelligent Systems and Computing, vol 1265, 166-174. doi:10.1007/978-3-030-58124-4_16

[3] Oriekhov, O, & Farionova, T. (2025). *A Five-Factor Nonlinear Regression Model for JAVA Applications Size Estimation,* CEUR Workshop Proceedings of the 6th International Workshop IT Project Management (ITPM 2025), Kyiv, Ukraine, May 22, 1-10. URL: https://ceur-ws.org/Vol-4023/paper1.pdf

[4] Olkin, I., & Sampson, A. R. (2001). *Multivariate Analysis: Overview*. in N. J. Smelser & P. B. Baltes, International encyclopedia of social & behavioral sciences (eds.) 1st edn., Elsevier, Pergamon, 10240–10247.

Oriekhov O. S., Farionova T. A.

Approaches to Reducing Multicollinearity Between Independent Factors in Nonlinear Regression Modelling for JAVA Application Assessment

Abstract. *The aim of this study is to reduce multicollinearity between independent factors by applying average metric values during the preprocessing of multivariate data, using the example of constructing a five-factor nonlinear regression model for processing code metrics of JAVA applications. It is demonstrated that the iterative outliers excluding from the empirical dataset increases VIF values and decreases model stability. The use of average values of the VMQ, TFQ, and CBO metrics relative to the total number of classes and interfaces made it possible to reduce the correlations between independent factors and avoid critical VIF levels.*

Keywords: *multicollinearity; VIF; nonlinear regression model; non-Gaussian data; JAVA-application.*