

УДК 519.21 : 519.24

ВИБІР ОПТИМАЛЬНОЇ МОДЕЛІ ЛІНІЙНОЇ РЕГРЕСІЇ ПРИ НАЯВНОСТІ СИЛЬНОЇ ЛІНІЙНОЇ ЗАЛЕЖНОСТІ МІЖ ВХІДНИМИ ЗМІННИМИ

В.В. Крохін¹, Д.П. Сівцов²

¹кандидат фізико-математичних наук, доцент кафедри ЕОМ, Дніпровський національний університет імені Олеся Гончара, м. Дніпро, Україна, e-mail: krokhinvolodymyr@yahoo.com

²завідувач навчальної лабораторії кафедри ЕОМ, Дніпровський національний університет імені Олеся Гончара, м. Дніпро, Україна, e-mail: explorer1024@ukr.net

Анотація. Проведено аналіз ефективності вибору оптимальної моделі множинної лінійної регресії (МЛР) при наявності сильної лінійної залежності між вхідними змінними. Дослідження проводилося за допомогою методу імітаційного моделювання. Проаналізовані п'ять методів вибору оптимальної МЛР (метод усіх можливих регресій з виправленим коефіцієнтом детермінації в якості критерію оптимальності; метод усіх можливих регресій з виправленим коефіцієнтом детермінації в якості критерію оптимальності і оцінкою значущості коефіцієнтів МЛР (на основі t-статистики); метод усіх можливих регресій з використанням статистики Маллоуза в якості критерію оптимальності; метод послідовного виключення; покроковий метод) з точки зору їх стійкості в умовах мультиколінеарності. Дослідження проводилося за допомогою спеціально розробленого в середовищі MATLAB програмного забезпечення.

Ключові слова: множинна лінійна регресія, оптимальна модель, критерій оптимальності, мультиколінеарності, імітаційне моделювання.

CHOICE OF THE LINEAR REGRESSION OPTIMAL MODEL IN THE PRESENCE OF STRONG LINEAR RELATIONSHIP BETWEEN INPUT VARIABLES

Volodimir Krokchin¹, Dmitro Sivtsov²

¹Ph.D. in Physics and Mathematical Sciences, Associate Professor, Department of Computer, Dnepr National University named by Oles Honchar, Dnepr, Ukraine, e-mail: krokhinvolodymyr@yahoo.com

²Head of the training laboratory, Department of Computer, Dnepr National University named by Oles Honchar, Dnepr, Ukraine, e-mail: explorer1024@ukr.net

Abstract. There has been undertaken an analysis of the effectiveness of choice of optimal model of multiple linear regression (MLR) with a strong linear relationship between the input variables. The study was conducted using a simulation method. It was conducted by analyzing five methods for selecting the optimal MLR (method of all possible regressions with the corrected coefficient of determination as an optimality criterion, the method of all possible regressions with the corrected coefficient of determination as benchmarks, and assessment of the significance of MDR ratios (based on t-statistics) method of all possible regressions

using statistics of Mallouza as an optimality criterion, successive elimination method, incremental method) in terms of their stability under conditions of multi-collinearity. The study was conducted with the help of specially designed tion in the environment MATLAB software.

Keywords: multiple linear regression, optimal model, optimality criterion, multicollinearity, simulation.

Вступ. Серйозною проблемою при оцінюванні коефіцієнтів в моделі множинної регресії є наявність лінійної залежності між деякими з вхідних змінних. [1,4,5]. Коли число незалежних змінних велике, "візуально" помітити наявність мультиколінеарності не можливо. На практиці слід говорити про можливу наявність квазімультиколінеарності між вхідними змінними. Інтерес представляє дослідження впливу мультиколінеарності на якість оцінювання коефіцієнтів моделі множинної лінійної регресії (МЛР). Зокрема, важливо виявити, як мультиколінеарність впливає на стійкість різних алгоритмів вибору оптимальної, в сенсі набору вхідних впливів, МЛР.

Мета роботи: Оцінити вплив ступеня лінійної залежності між вхідними змінними на стійкість різних алгоритмів вибору оптимальної моделі лінійної регресії (МЛР) шляхом імітаційного моделювання.

Матеріали та результати дослідження: Розглянемо модель множинної лінійної регресії (МЛР):

$$Y = X\beta + U.$$

У цій моделі Y – спостережуваний $[n \times 1]$ вектор значень залежної змінної; X – $[n \times k]$ матриця, у якій перший стовпець складається з одиниць, а інші стовпці відображають спостережувані значення залежних змінних; β – невідомий $[k \times 1]$ вектор параметрів (коефіцієнтів) моделі; U – неспостережуваний $[n \times 1]$ вектор збурень (помилки моделі).

Якщо стовпчики матриці X є лінійно незалежними, жоден стовпець не може бути представлений як лінійна комбінація інших стовпців цієї матриці. Наявність мултіколінеарності означає, що детермінант матриці $(X'X)$ дорівнює нулю, тобто мултіколінеарність означає, що детермінант $(X'X) = 0$. Однак точна коллінеарність між даними зустрічається рідко. Тому слід очікувати, що між даними, що використовуються як незалежні змінні можемо мати:

$$\lambda_1 X_1 + \lambda_2 X_2 + \dots + \lambda_k X_k + v = \vec{0},$$

де не всі $\lambda_j = 0$, а v – випадкова величина.

Якщо стовпці матриці X пов'язані таким співвідношенням то між незалежними змінними існує квазіколінеарність. Її наявність призводить до

того, що стандартні помилки деяких оцінок коефіцієнтів лінійної регресії стають дуже великими [1], тому невеликі зміни у вихідних даних, призводять до великих змін оцінок коефіцієнтів регресії.

Нами було досліджено вплив квазіколінеарності на стійкість методів вибору оптимальної МЛР.

Для вибору оптимальної МЛР застосовувалися 5 критеріїв [2,3]:

- метод усіх можливих регресій з виправленим коефіцієнтом детермінації в якості критерію оптимальності (MR2);
- метод усіх можливих регресій з виправленим коефіцієнтом детермінації в якості критерію оптимальності і оцінкою значущості коефіцієнтів МЛР (на основі t -статистики) (MR2t);
- метод усіх можливих регресій з використанням статистики Маллоуза в якості критерію оптимальності (Mlz);
- метод послідовного виключення (BWE);
- покроковий метод (SWP).

Ефективність цих методів при наявності квазімультіколінеарності серед вхідних змінних досліджувалася за допомогою імітаційного моделювання. На ЕОМ моделювалося m випадкових послідовностей, які використовувалися в якості можливих значень незалежних змінних. З них вибиралися k послідовностей в якості значень змінних, що беруть участь в побудові МЛР. Потім передбачалося, що невідомо які і скільки з m незалежних змінних входять в справжню модель. Таким чином, пошук моделі здійснювався з використанням різних можливих комбінацій з m незалежних змінних на основі перерахованих вище критеріїв. Вказана методика була реалізована на платформі MATLAB. Так, було проведено дослідження впливу мультиколінеарності вхідних змінних на якість ідентифікації оптимальної моделі. Експерименти проводилися при таких значеннях параметрів імітаційного моделювання:

- кількість можливих вхідних змінних $m = 10$;
- кількість вхідних змінних використаних для побудови МЛР $k = 5$;
- обсяг вибірки – 50;
- кількість експериментів – 500.

Вдалою ідентифікацією вважається така, що вибір МЛР, в яку в якості вхідних змінних включені в повному обсязі тільки такі, які використовувалися при моделюванні регресійної залежності [1,2]. З іншого боку, помилки при ідентифікації МЛР можливі такі [4]:

- не враховані деякі з "справжніх" вхідних змінних;
- враховані певні "зайві" вхідні змінні.

Експериментальні дослідження дозволили виявити, що:

При наявності сильної мультиколінеарності між двома вхідними змінними, одна з них несе майже повну інформацію про іншу. Звідсіля витікає, що не включення в модель одночасно обох змінних призводить до серйозної помилки ідентифікації.

Покроковий метод (SWP) вибору оптимальної МЛР забезпечує найвищий відсоток правильних ідентифікацій при лінійної залежності між двома вхідними змінними і коли кореляція між ними $\rightarrow 1$.

При кореляції $< 0,995$ повністю правильну ідентифікацію забезпечує метод усіх можливих регресій на основі статистики Маллоуза, як критерію оптимальності.

Найбільш швидким є метод послідовного виключення (BWE), який забезпечує правильні ідентифікації МЛР на рівні $> 90\%$ при кореляції між вхідними змінними $< 0,99995$.

Висновки. Числові данні, отримані в результаті імітаційного моделювання дозволяють стверджувати, що суттєвий вплив квазіколінеарності виявляється тільки при високий ступені корельованості між вхідними змінними, а саме, $> 0,999$. Якщо серед незалежних змінних є сильна лінійна залежність (квазіколінеарність), то для вибору оптимальної МЛР слід застосовувати метод усіх можливих регресій зі статистикою Маллоуза в якості критерію оптимальності, якщо число можливих незалежних змінних ≤ 15 . В іншому випадку вигідніше використовувати метод послідовного виключення.

ЛІТЕРАТУРА

1. Дрейпер Н., Сміт Г. - Прикладний регресійний аналіз: У 2-х кн. 2 / Пер. з англ. - 3-е изд., перераб. і доп. - М.: Діалектика. 2016. - 912с.
2. Крохін В. В. Алгоритми ідентифікації багатопараметричних регресійних моделей // Навчальний посібник до Вивчення курсу «Цифрова обробка сигналів». - Дніпропетровськ.: РВВ ДНУ, 2012. - 32 с.
3. Крохін В. В., Кузьменко Н. О. Автоматизація вибору оптимальної моделі лінійної регресії // Системні технології. Регіональний міжвузівський збірник наукових праць. - Випуск 1 (78). Дніпропетровськ, 2012 - з. 73 - 83.
4. Gujarati D. N. Econometría. Tercera edición. McGraw-Hill - Santafe de Bogotá, 1997. - 824 p.
5. Rawlings J.O., Pantula S.G., Dickey D.A. Applied Regression Analysis. A Research Tool. Second edition - New York: Springer, 2001. - 671 p.