УДК 330.43(075.8)

ПРОБЛЕМА МУЛЬТИКОЛІНЕАРНОСТІ В ЗАДАЧІ ЛІНІЙНОЇ РЕГРЕСІЇ: ЕКОНОМІЧНИЙ ТА МАТЕМАТИЧНИЙ АСПЕКТИ

Тижненко О. Г., к.ф.-м.н., доцент, ХНЕУ ім. С. Кузнеця, Харків, Україна

Економічний зміст задачі багатофакторної лінійної регресі міститься у визначенні значимої оцінки впливу кожного з регресорів на відгук за умовою сталості усіх інших регресорів. Ця задача розглядається в даної роботі за припущенням про нормальний розподіл випадкової помилки, $\varepsilon \sim \mathbb{N}(0, \sigma \mathbf{I})$ та припущенням, що в ГС співвідношення між регресорами лінійні. Змінні в ГС вважаються стохастичними та нормально розподіленими. Таку ГС будемо називати лінійною та нормально розподіленою, ЛНГС.

З економічної, а взагалі, з фізичної точки зору задача лінійної регресії полягає: 1) у формулюванні моделі регресії адекватної даним в ЛНГС та 2) визначенні оцінок усередненого впливу регресорів на відгук, близьких до їх значень в ЛНГС.

З математичної точки зору, якщо застосовується МНК, то вирішується задача найкращого наближення регресії до відгуку [1, 2, 3], де:

1) Задача лінійної регресії формулюється як задача узгодження спостережених та розрахункових даних;

2) За умовами теореми Гауса-Маркова теоретично доводиться, що МНК-рішення є BLUE (при цьому не виключається, що існують *вибіркові оцінки*, які зроблені іншими методами, з меншою дисперсією, чим дає МНК, але вони не є незміщеними);

3) Використовуються теоретичні формули для оцінок дисперсій компонент МНК-рішення для вибіркових даних будь-якого розміру, що дозволяє, за додатковою умовою нормальності випадкової помилки, зробити відповідні статистичні інференції щодо оцінок вибіркових коефіцієнтів регресії.

4) Доводиться асимптотична спроможність МНК-рішення (незалежно від наявності мультиколінеарності).

Основною перешкодою, яка заважає МНК-рішенню бути адекватним економічному рішенню задачі лінійної регресії в ЛНГС для не дуже великих вибірок є майже-колінеарність даних [4], яка призводить неприйнятну варіабельність МНК-рішень. Тобто, з точки зору рішення математичної проблеми узгодження регресії з відгуком, МНК чудово справляється зі своєю задачею при будь-якої степені майже-колінеарності спостережених даних, але зі зростанням рівня майже-колінеарності даних катастрофічно зростає дисперсія МНК-рішення і тим більше, чим менше розмір вибірки.

Іншим джерелом помилок рішення задачі лінійної регресії незалежно від методу її рішення є нелінійність ГС, яка досліджується. Ця проблема стосується неадекватності моделі і може бути виключена, в принципі, за допомогою перетворення даних на етапі препідготовки даних.

Дуже важливим джерелом помилок рішення задачі лінійної регресії незалежно від методу її рішення є невірна специфікація моделі, але ця проблема стосується тільки економічного аспекту дослідження і в даної роботі не розглядається.

Різним способам боротьби з проблемою майже-колінеарності присвячено безліч робот, огляд яких даний, наприклад, в [4-7]. В основному вони присвячені методам діагностики існування занадто тісної взаємної кореляції між регресорами з ціллю зменшення рівня майже-колінеарності шляхом відбору регресорів. Цій шлях не приводе, однак, до позитивного результату, оскільки рівень коліг-

неарності змінюється неперервно і не має критичних точок [5, 8-11].

Що стосується впливу майже-колінеарності даних на варіабельність МНК-рішення, то попередні дослідження однозначно показали необхідність створення нових методів рішення економічної задачі регресії, які б давали прийнятну з економічної точки зору варіабельність рішення при невеликих розмірах вибірки зі середнім значенням рішення близьким до значень в ГС.

У даної роботі запропоновано метод рішення задачі лінійної регресії, який дає рішення близькі до коефіцієнтів регресії в ГС з прийнятною варіабельністю для не тільки для дуже великих вибірок, як МНК, але й для середніх та малих вибірок при будь-якому рівні майже-колінеарності економічних даних.

Цей метод являє собою модифікований метод Крамера(ММК) рішення регуляризованого МНК-рівняння

$$(X'X + \alpha I)b = X'Y$$

при $\alpha \ll 1$. На відміну від ridge regression, метод не потребує підбору постійної регуляризації α для кожного рішення і дає практично незміщене рішення з малою дисперсією при $\alpha = 0.01$ при будь-якому рівні майже-колінеарності даних.

Для рішення погано обумовлених матричних рівнянь, замість методу Гауса або Крамера пропонується метод (ММК), який дає рішення матричного рівняння з суттєво меншою варіабельністю, ніж метод Крамера або Гауса. Як і метод рідж-регресії, ММК є наближеним методом, але на відміну від нього, дає практично нульове зміщення. ММК використовує наступний метод рішення стандартизованого матричного МНК-рівняння (1)

$$Ax = B\ ,$$

яке записується у вигляді:

$$A'Ax = A'B\ .$$

Позначимо:

$$A'A = H_1,\ A'B = B_1.$$

Тобто рішаємо рівняння:

$$H_1 x = B_1$$

Враховуючи можливу погану обумовленість матриці H_1, зменшуємо число обумовленості замінюючи матрицю H_1 на близьку матрицю

$H = H_1 + \alpha E,$

де E є одиничною матрицею розміру матриці H_1, та $0 < \alpha \ll 1$ (оптимальне значення $\alpha = 0.001$). Згідно з формулою Крамера, рішення рівняння $Hx = B_1$ запишемо:

$$x_j = \frac{\Delta_j}{\Delta}, \qquad (1)$$

$$\Delta_j = \sum_{k=1}^{n} (-1)^{j+k} B_1(k) \det\left(H(t_k, t_j)\right)$$

$$\Delta = \sum_{k=1}^{n} (-1)^{j+k} H(k,j) \det(H(t_k, t_j))$$

та $t_k = 1,2,\dots,k-1,k+1,\dots,n$. Тут $H(t_k, t_j)$ є матрицею H, з котрої викреслені k-й рядок та j-й стовпчик. $H(k,j)$ є елемент матриці H. Тобто формули (1) є звичайними формулами розкладу визначника за алгебраїчними доповненнями.

Ми можемо помножити чисельник та знаменник у формулі (1) на будь-якій визначник не рівний нулю розміру матриці $H(t_k, t_j)$. Позначимо як H_j^{-1} обернену матрицю до матриці H, у котрої викреслено діагональний елемент $H(j,j)$, тобто j-й рядок та j-й стовпчик. Для кожного j у формулі (4) помножимо чисельник та знаменник на $\det\left(H_j^{-1}\right)$.

Використовуючи властивість визначників:

$$\det(AB) = \det(A)\det(B)\ ,$$

ми можемо записати рішення (1) в іншому вигляді:

$$\Delta_j = \sum_{k=1}^{n} (-1)^{j+k} B_1(k) \det\left(H_j^{-1} H(t_k, t_j)\right)$$

$$\Delta = \sum_{k=1}^{n} (-1)^{j+k} H(k,j) \det(H_j^{-1} H(t_k, t_j))$$

Наближене рішення СЛАР (1) тоді запишеться як

$$x_j = \frac{\Delta_j}{\Delta}.$$

Як показали дослідження, ця проста операція призводить до суттєвої стабілізації рішення погано-обумовлених матричних рівнянь при оптимальному значенні $\alpha = 0.001$.

Недоліком ММК є необхідність обчислювати багато разів визначники розмірності числа регресорів. Це, по перше, призводить до значних обчислювальних витрат, а по друге, до накопичування обчислювальних помилок при обчислюванні визначників значних порядків, що запобігає отриманню коректного рішення.

Слід однак відмітити, що завдяки центруванню даних обчислювальні помилки при обчислюванні визначників взаємно компенсуються, що дає можливість вирішувати з прийнятною точністю задачі регресії з числом регресорів близько ста и вище.

Програма рішення стандартизованого МНК-рівняння в R:

```
kramer <- function(A, B, alpha=1e-3) {
    n <- length(B)
  H1 <- t(A)%*%A
  B1 <- t(A)%*%B
  E <- diag(n)
  H2 <- alpha*E+H1
  row_idx <- sort(rep(seq_len(n), n))
  col_idx <- rep(seq_len(n), n)
  D <- vector("double", n)
  D1 <- vector("double", n)
  for(item in seq_len(n*n)) {
   i <- row_idx[item]
   k <- col_idx[item]
    if (k == 1) {
    H3 <- solve(H2[-i, -i], diag(n-1))
   }
   detH3H2 <- det(H3%*%H2[-k, -i])
   D[i] <- D[i] + (-1)^(i+k)*B1[k]*detH3H2
   D1[i] <- D1[i] + (-1)^(i+k)*H2[k,
i]*detH3H2
  }
  X <- D/D1
  cbind(X)
 }
```

Список використаної літератури

1. Adkins L. C., Hill R. C. (2001). Collinearity. Companion in Theoretical Econometrics, edited by Badi Baltagi. Oxford: Blackwell Publishers, Ltd., pp. 256-278.
2. Seber, G.A.F., and Lee, A.J. (2003). Linear Regression Analysis, 2nd edition, Wiley, NY.
3. Wooldridge, J.M. (2009), Introductory Econometrics: Modern Approach, 5th ed. (South-Western: Ohio).
4. Adkins L. C., M. S. Waters, R. C. Hill, (2015). "Collinearity Diagnostics in gretl," Economics Working Paper Series 1506, Oklahoma State University, Department of Economics and Legal Studies in Business.
5. Harvey, A.C. 1977. "Some Comments on Multicollinearity in Regression." Applied Statistics 26(2): 188-191.
6. Badi Baltagi (2011). Econometrics. Springer.
7. Hill R. C., L. C. Adkins. (2003) Collinearity, A Companion to Theoretical Econometrics, Edited by Badi H. Baltagi: Blackwell Publishing Ltd, p. 256-278.
8. Marquardt, D.V. (1970). "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation", Technometrics, 12: 591-612
9. Spanos, A. and A. McGuirk (2002). "The Problem of Near-Multicollinearity
Revisited: erratic vs. systematic volatility," Journal of Econometrics, 108: 365-393.
10. Wooldridge, J.M. (2009), Introductory Econometrics: Modern Approach, 5th ed. (South-Western: Ohio).
11. Rao, C. R., H. Toutenberg (1999), "Linear Models: Least Squares and Alternatives" 2nd ed., Springer.

Автор:

Тижненко Александр Григорович, доцент, ХНЕУ ім. С. Кузнеця.

oleksandr.tyzhnenko@m.hneu.edu.ua

Тези доповіді надійшли 09 лютого 2018 року.

Опубліковано в авторській редакції.