УДК 330.43(075.8)

ДІАГНОСТИКА МУЛЬТИКОЛІНЕАРНОСТІ ТА ЗАСТОСУВАННЯ RIDGE-РЕГРЕСІЇ В R

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

Железнякова Е. Ю., к.ф.-м.н., доцент, ХНЕУ ім. С. Кузнеця, Харків, Україна

***Анотація* – У роботі розглянуті особливості застосування МНК-рішення лінійної регресійної моделі в умовах наявності мультиколінеарності. Розглянуті властивості МНК-рішення при наявності мультиколінеарності. Розглянуті також існуючи методи діагностики мультиколінеарності та запропоновано новий метод діагностики. Визначені умови коректного застосування ridge-регресії при наявності мультиколінеарності.**

***Ключові слова* – мультиколінеарність, ridge-регресія, VIF-фактор, МНК, обумовленість, сингулярні числа.**

Основною проблемою застосування МНК-рішення в економічних дослідженнях є мультиколінеарність даних [1, 2], яка приводить до поганої обумовленості МНК-матриці. Інтенсивність мультиколінеарності виміряють, як правило, за допомогою VIF-фактору. Ступінь поганої обумовленості МНК-матриці виміряють числом обумовленості, *cond* , яке обчислюється як відношення мінімального сингулярного числа МНК-матриці до її максимального сингулярного числа. За критерієм VIF-фактора вважається, що має місце мультиколінеарність, якщо $VIF > 4$. З точки зору обумовленості МНК-матриці вважається, що МНК-матриця є погано обумовленою, якщо її число обумовленості $cond > 10$ [1].

Показники мультиколінеарності, VIF-фактор та число обумовленості, *cond* , в R обчислюються за допомогою функцій vif() та kappa() [3]. Якщо X є матриця регресорів, яка складається з регресорів X_1, X_2 та X_3, а Y є відгук [2], то для застосування функції vif() необхідно викликати проект car:

```
> library(car)
> fit <- lm(Y~X1+X2+X3, Egor9)
> vif(fit)
      X1       X2       X3
116.01489 151.66089  68.74305
```

Серед чисел необхідно обрати максимальне. Для розрахунку числа обумовленості МНК-матриці необхідно побудувати дизайн-матрицю [1] даних $D = [1\ X]$. Тоді число обумовленості МНК-матриці, cond = = kappa(D'D).

Загальновідомо, що при наявності мультиколінеарності рішення задачі лінійної регресії може бути некоректним у тому сенсі, що отримані оцінки параметрів лінійної регресійної моделі ($\hat{b}_j$) можуть дуже сильно відрізняться від відповідних значень у генеральної сукупності (b_j). Зі статистичної точки зору, це є результатом дуже великої дисперсії ($s_{\hat{b}_j}$) оцінок параметрів лінійної регресійної моделі ($\hat{b}_j$), яка визначається величиною стандартного відхилення залишкової помилки у генеральній сукупності (σ_e) та діагональними елементами оберненої МНК-матриці [1]:

$$s_{\hat{b}_j} = \sigma_e \sqrt{C_{jj}},\ C_{jj} = diag(X'X)^{-1}_{jj}, \qquad (1)$$

Слід зауважити, що стандартне відхилення залишкової помилки σ_e в МНК-методі є невеликою величиною і не зростає при збільшенні степені мультиколінеарності. Однак це не стосується величин C_{jj}. Саме ці величини суттєво зростають при збільшенні VIF-фактора або числа обумовленості ($cond$), що призводить до неприйнятності МНК-рішення, бо при цьому коефіцієнти регресії стають некоректними завдяки великої помилки рішення.

Далі слід відзначити дуже важливий факт: умови наявності мультиколінеарності [1]:

$$VIF > 4,\ cond > 10, \qquad (2)$$

часто є занадто суворими, що може привести до необґрунтованої втрати довіри к даним. Справа у тому, як визначати статистичну незалежність регресорів, оскільки вважається, що для незалежних регресорів $VIF = 1$ та $cond = 1$.

Слід відмітити, що в економічних даних незалежних регресорів не буває взагалі, якщо їх не створювати навмисно. У практиці моделювання даних незалежними вважаються дані генераторів псевдовипадкових чисел (ГПЧ). В R, наприклад, такі числа для стандартного нормального розподілу генеруються функцією rnorm() [3].

Регресори, які генеруються за допомогою ГПЧ, слід вважати статистично незалежними.

Це дає підставу для більш точної діагностики наявності мультиколінеарності. Якщо підставити у матрицю регресорів $X = (X_j)$ замість самих регресорів X_j псевдовипадкові вектори R_j та обчислити діагностичні фактори $VIF_R(X)$ та $cond_R(D'D)$ для дизайн-матриці, то слід вважати, що мультиколінеарність існує, якщо спостережені значення цих факторів, VIF та $cond$ перевищують критичні значення, VIF_R та $cond_R$:

$$VIF > VIF_R,\ cond > cond_R. \qquad (3)$$

Слід відмітити, що критичні значення VIF_R та $cond_R$ можуть суттєво перевищувати одиницю і це залежить від відношення m/n, де n - розмір вибірки, а m - число регресорів.

Якщо, наприклад, $n = 10, m = 5$, то критичні значення: $VIF_R = 2,58$, $cond_R = 16$. Тобто число обумовленості перевищує 10 – загально прийняте критичне значення, але мультиколінеарність ще не спостерігається. У даному разі мультиколінеарність виникає при $cond_R > 16$. Тобто, якщо число обумовленості МНК-матриці для зазначених даних близьке до 20, то не слід очікувати мультиколінеарності, незважаючи на загальноприйнятну умову (2). Це означає, що для цих даних ми можемо довіряти результатам розрахунків коефіцієнтів регресії та відповідним статистичним інференціям і використовувати їх і економічних дослідженнях.

Тобто умова (3) дозволяє точніше враховувати наявність мультиколінеарності у спостережених даних, що, у свою чергу, дозволяє використовувати МНК для тих економічних даних, для яких умова (2) цього не дозволяє.

Незважаючи на таку можливість, реальні економічні дані такі, що спостережені показники мультиколінеарності VIF та $cond$ суттєво перевищують навіть критичні значення VIF_R та $cond_R$, що не дозволяє довіряти МНК-рішенню задачі регресії.

У цьому разі немає іншого виходу, як використати ridge-регресію [1, 2]. Однак, слід мати на увазі, що застосування ridge-регресії неможливе для лінійної моделі регресії у натуральних змінних:

$$Y = b_0 + \sum_{j=1}^{m} b_j X_j + e, \qquad (4)$$

яка використовується практично у всіх підручниках та пакетах прикладних програм.

Зауважимо, що МНК-матриця $D'D$ у натуральних змінних з дизайн-матрицею D має погані характеристики: дуже великі значення числа обумовленості та максимального сингулярного числа. У ridge-регресії замість матриці $D'D$ використовується матриця регуляризації: $D'D + ridge\, I$, де ridge – параметр ridge-регресії, а I – одинична матриця. Суть ridge-регресії полягає в тому, що матриця регуляризації має сингулярні числа на величину ridge більшу, ніж матриця $D'D$.

Якщо максимальне сингулярне число матриці $D'D$ дуже велике, то додавання малого параметру ridge несуттєво змінює число обумовленості матриці регуляризації $D'D + ridge\, I$, яке визначається як відношення максимального сингулярного числа до мінімального. Це означає, що матриця регуляризації у ridge-методі буде, як і раніше, погано обумовленою.

Таким чином, у ridge-методі використовується тільки лінійна регресійна модель у стандартизованих змінних.

$$t_Y = \sum_{j=1}^{m} \beta_j t_j + e'. \qquad (5)$$

Слід також зауважити, що ridge-метод не завжди може бути застосованим взагалі, якщо МНК-матриця $(t't)$ регресійної моделі у стандартизованих змінних погано обумовлена за рахунок досить великого максимального сингулярного числа при будь-якому мінімальному сингулярному числі. У цьому разі додавання малого параметру ridge несуттєво змінює число обумовленості матриці і вона залишається погано обумовленою.

Таким чином, перш ніж вирішувати задачу лінійної регресії необхідно розрахувати: критичні та спостережувані значення VIF та $cond$, VIF_R та $cond_R$, а також мінімальні та максимальні сингулярні числа МНК-матриці для стандартизованої форми моделі, та вирішити, чи можливо взагалі знайти прийнятне для економічного аналізу рішення.

Якщо спостережувані дані дають $VIF > 4$, то їх слід трансформувати до стандартизованого виду. Для стандартизованого виду характерне суттєве зменшення як VIF -фактору, так і числа обумовленості МНК-матриці, що для коректності МНК-рішення не має ніякого значення: обидва матричних МНК-рівняння одночасно або мають, або не мають рішенн. Однак МНК-матриці обох рівнянь мають дуже різні сингулярні числа, що дозволяє застосовувати ridge-регресію якщо мінімальне сингулярне число МНК-матриці стандартизованої моделі лінійної регресії набагато менше одиниці, а максимальне сингулярне число не набагато більше одиниці.

Прийнятним критерієм неможливості застосування ridge-регресії є сталість рішення задачі регресії при зміненні ridge-параметру. Якщо при збільшенні ridge-параметру від 10 % значення мінімального сингулярного числа стандартизованої МНК-матриці до одиниці рішення ridge-регресії практично не відрізняється від МНК-рішення, то метод ridge-регресії не може бути застосованим. Це означає, що задача лінійної регресії не може бути вирішеною у рамках МНК-наближення.

Список використаної літератури

1. Greene W. H. Econometris Anakysis. – Pearson: Prentice Hall, 2012. – 1239 p.

2. Егоршин А. А. Корреляционно-регрессионный анализ / А. А. Егоршин, Л. М. Малярец. – Х: «Основа», 1998, - 208 с.

3. Kleiber C. Applied Econometrics with R / C. Kleiber, A. Zeiles: Springer, 2008, - 229 p.

Автори

Тижненко Олександр Григорович, доцент, ХНЕУ ім. С. Кузнеця, olexsandr.tyzhnenko@m.hneu.edu.ua.

Железнякова Еліна Юріївна, доцент, ХНЕУ ім. С. Кузнеця, elina.zhelezniakova@m.hneu.edu.ua

Тези доповіді надійшли 05 лютого 2017 року.

Опубліковано в авторської редакції.