ТЕМА 3
Мультиколінеарність та її вплив на оцінки параметрів моделі

У процесі створення моделей дослідник стикається із значною кількістю перешкод, що викликаються специфікою включених даних, обраною формою функціональної залежності між змінними, економічною гіпотезою тощо. Однією з таких проблем є мультиколінеарність, наявність якої заважає отримати достовірні результати моделювання та здійснювати аналіз функціональних зв'язків. У темі 3 розглянуто наслідки мультиколінеарності та наведено способи її тестування й вилучення.

Основні питання, що розглядаються:

1. Визначення мультиколінеарності та її природа
2. Теоретичні та практичні наслідки мультиколінеарності в загальному випадку
3. Тестування мультиколінеарності та засоби її вилучення

Основні терміни:

Мультиколінеарність; BLUE-оцінки; незміщена оцінка; ефективні оцінки; характеристичні значення; дисперсійно-інфляційний VIF-фактор; рівняння перших різниць; факторний аналіз; метод головних компонент; гребенева регресія.

3.1. ВИЗНАЧЕННЯ МУЛЬТИКОЛІНЕАРНОСТІ ТА ЇЇ ПРИРОДА

Мультиколінеарність виникає тоді, коли більше, ніж два фактори зв'язані між собою лінійною залежністю, тобто має місце вплив факторів один на одного. Наявність мультиколінеарності буде означати, що деякі фактори завжди будуть діяти в унісон. Іншими словами, коефіцієнт коре-ляції між цими двома факторами має значення, близьке або дорівнює 1.

Наприклад, мультиколінеарність може бути проблемою, коли ми вивчаємо залежність між ціною акції, дивідендами на акцію та зароблением прибутком на акцію, оскільки дивіденди та зароблений прибуток на одну акцію мають високий ступінь кореляції.

Мультиколінеарність може виникати за різних умов:

1. Є глобальна тенденція до одночасної зміни економічних показників. На макроекономічні показники впливають однакові фактори. Це призводить до того, що вони відображають широкий спектр моделей однакової економічної ситуації. Наприклад, у періоди спаду однаково можуть спадати показники доходу та споживання, інвестицій та зайнятості тощо. Сама наявність трендів у динамічних рядах є причиною мультиколінеарності.
2. Широке використання в економетричних моделях лагових значень однієї змінної також призводить до виникнення мультиколі-неарності. Наприклад, добре відомі інвестиційні функції, в яких лагові значення минулого рівня економічної активності вводяться як окремі змінні. У функціях споживання витрати на споживання у попередньому періоді вводяться в модель поряд з величиною поточного рівня доходу.

Які труднощі викликає мультиколінеарність у процесі моделювання та аналізу моделей? В результаті варіація у вихідних даних припиняє бути повністю незалежною і не можна досліджувати вплив кожного фактору окремо. Чим сильніше мультиколінеарність факторів, тим менш надійна оцінка розподілу суми поясненої варіації за окремими факторами за допомогою методу найменших квадратів.

Чому для класичної лінійної моделі вимогою одного з припущень є відсутність мультиколінеарності між її факторами? Тому що: а) параметри регресії стають невизначеними, тобто якщо припустити, що $x_1=a\cdot x_2$, то при певних перетвореннях ми отримаємо, що параметр $b_1=0/0$, $b_2=0/0$, б) а їхні середні квадратичні відхилення прямують до нескінченності, тобто якщо $x_1=kx_2$, $r(x_1x_2)=1$, то маємо, що

$$var(b_1)=\sigma^2_e / (1-r^2)\Sigma(x_{1i}- \bar{x}_1)^2$$

$$var(b_2)=\sigma^2_e / (1-r^2)\Sigma(x_{2i}- \bar{x}_2)^2,$$

тоді $var(b_1)=\infty$, $var(b_2)=\infty$. Відповідно, дорівнюють нескінченності і середньоквадратичні відхилення.

Звичайно, досконала мультиколінеарність є дуже рідкісним явищем; частіше в економічних дослідженнях немає точної лінійної залежності між параметрами.

3.2. ТЕОРЕТИЧНІ ТА ПРАКТИЧНІ НАСЛІДКИ МУЛЬТИКОЛІНЕАРНОСТІ В ЗАГАЛЬНОМУ ВИПАДКУ

Якщо умови класичної моделі задовольняються, то МНК-оцінки (оцінки, обчислені за методом найменших квадратів) є BLUE (best linear unbiased estimator – найкраща лінійна оцінка без відхилень). ***BLUE-оцінка*** означає, що ці оцінки лінійні, без відхилень, мають найменшу дисперсію з усіх можливих методів оцінювання. Наприклад, при значеннях $y_1, y_2, \ldots, y_n$ лінійна оцінка має вигляд: $k_1y_1+k_2y_2+\ldots+k_ny_n$, де k_i, i = 1, n – константи.

Навіть при дуже високій, але недосконалій мультиколінеарності МНК-оцінки все ще зберігають властивість BLUE-оцінок, тому мультиколінеарність стає причиною таких теоретичних наслідків:

– у разі високої мультиколінеарності МНК-оцінки є незміщеними.

Незміщеність оцінок означає, що математичне сподівання випадкової величини дорівнює нулеві. Це означає, що при великій кількості вибіркових оцінюваних залишки не будуть накопичуватися і знайдений параметр регресії b_i можна розглядати як середнє значення з можливої великої кількості незміщених оцінок;

– ***колінеарність*** не порушує властивостей мінімуму дисперсії: в класі лінійних незміщених оцінок МНК-оцінки мають мінімальну дисперсію, і тому вони ефективні. Однак це не означає, що дисперсія МНК-оцінки буде неминуче малою (відносно значення параметра) в будь-якій з наведених вибірок. Оцінки називаються ***ефективними***, якщо вони характеризуються найменшою дисперсією (виникає можливість переходу від точкової оцінки до інтервальної);

– мультиколінеарність – це явище виключно регресійного аналізу вибірки в тому розумінні, що навіть якщо змінні *x* пов'язані в генеральні сукупності нелінійно, вони можуть мати лінійний зв'язок в кожному окремому випадку: коли ми постулюємо вибіркову або узагальнену функцію регресії, то впевнені, що всі змінні величини *x*, які утворюють модель, мають окремий або незалежний вплив на залежну змінну величину *y*. Однак може так трапитися, що в будь-якій з вибірок, яка використовується для перевірки узагальненої моделі, деякі або всі змінні величини *x* настільки високо колінеарні, що ми не можемо виявити їхнього індивідуального впливу на *y*. Наша вибірка, так би мовити, нас підводить, хоча за теорією всі *x* важливі. Іншими словами, вже неможливо застосовувати всі змінні величини *x* в аналізі.

Незважаючи на те, що МНК-оцінки при мультиколінеарності є повністю BLUE-оцінками, вона має достатньо негативні практичні наслідки для моделювання.

Першим практичним наслідком мультиколінеарності є велика дисперсія і коваріація оцінок параметрів, обчислених за методом найменших квадратів.

$$\mathrm{var}(b_1) = \frac{\sigma_\varepsilon^2}{(1-r^2)\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2};$$

$$\mathrm{var}(b_2) = \frac{\sigma_\varepsilon^2}{(1-r^2)\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)^2};$$

$$\mathrm{cov}(b_1, b_2) = \frac{-r\sigma_\varepsilon^2}{(1-r^2)\sqrt{\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2 \cdot \sum_{i=1}^{n}(x_{2i}-\bar{x}_2)^2}},$$

де *r* – коефіцієнт кореляції між x_1 і x_2.

Другим практичним наслідком мультиколінеарності є збільшення інтервалу довіри. Оскільки збільшення коефіцієнту кореляції призводить до збільшення значень середньоквадратичних відхилень параметрів, то, звичайно, збільшується й інтервал довіри до них (див. табл. 3.1).

Таблиця. 3.1

Ефект впливу збільшення коефіцієнту кореляції на інтервал довіри для параметра β_1

Значення r_{x1x2}	**95 % інтервал довіри для β_1**
0	$b_1 \pm 1{,}96\sqrt{p}$
0,5	$b_1 \pm 1{,}96\sqrt{1{,}33 \cdot p}$
0,95	$b_1 \pm 1{,}96\sqrt{10{,}26 \cdot p}$
0,999	$b_1 \pm 1 = 1{,}96\sqrt{500 \cdot p}$

Третім практичним наслідком мультиколінеарності є незначущість *t*-статистики. У випадку мультиколінеарності $\mathcal{E}_{b_1}$ нескінченно зростає, а *t*-значення прямує до нуля, оскільки:

$$t = \frac{b_1}{\mathcal{E}_{b_1}}.$$

Мультиколінеарність *не є проблемою*, коли єдиною метою регре-сійного аналізу є прогноз, оскільки чим вище значення коефіцієнту детермінації R^2, тим точніший прогноз. Це справедливо доти, доки значення залежних змінних, для яких і здійснюється прогноз, мають однакову майже лінійну залежність з початковою матрицею *X*. Таким чином, якщо у побудованій регресії встановлено, що приблизно $x_1 \approx 2 \cdot x_2$, то в наступних прикладах прогнозування x_1 має бути приблизно рівним $2 \cdot x_2$. Ця умова майже нездійсненна на практиці.

Крім того, якщо метою побудови множинної регресії є не прогноз, а виявлення впливу за допомогою коефіцієнтів еластичності результа-тивної ознаки *y* за кожним з факторів, мультиколінеарність перетворюється на проблему.

Також мультиколінеарність не загрожує якості моделі, якщо R^2 великий і параметри регресії є значущими, оскільки *t*-статистика висока.

3.3. ТЕСТУВАННЯ МУЛЬТИКОЛІНЕАРНОСТІ ТА ЗАСОБИ ЇЇ ВИЛУЧЕННЯ

Єдиної мети для визначення мультиколінеарності немає. Наведемо кілька методів тестування наявності мультиколінеарності.

1. Високе значення R^2 і незначущість *t*-статистики. Одночасна наявність цих двох факторів є «класичною» ознакою мульти-колінеарності.

2. Високе значення парних коефіцієнтів кореляції. Якщо значення хоча б одного з парних коефіцієнтів кореляції (між змінними x_i) більше 0,8, то мультиколінеарність є серйозною проблемою. Ця умова є необхідною, проте недостатньою, оскільки мультиколінеарність може бути навіть при невеликих значеннях парних коефіцієнтів кореляції у більш, ніж двофакторній регресійній моделі. Для перевірки цього тесту будується матриця кореляції *R*:

$$R = \begin{pmatrix} & \overline{y} & \overline{x_1} & \overline{x_2} & \overline{\dots} & \overline{x_k} \\ \overline{y} & r_{y^2} & r_{yx_1} & r_{yx_2} & \dots & r_{yx_k} \\ x_1 & r_{yx_1} & r_{x_1^2} & r_{x_1x_2} & \dots & r_{x_1x_k} \\ x_2 & r_{yx_2} & r_{x_1x_2} & r_{x_2^2} & \dots & r_{x_2x_k} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_k & r_{yx_k} & r_{x_1x_k} & \dots & \dots & r_{x_k^2} \end{pmatrix}$$

3. F-тест для визначення мультиколінеарності. Цей тест було запропоновано Глаубером і Фарром. Наявність мультиколінеарності свідчить про те, що один або більше факторів пов'язані між собою лінійною або приблизно лінійною залежністю. Одним із способів визначення щільності регресійного зв'язку є побудова регресійної залежності кожного фактора x_i з усіма іншими факторами і обчислення відповідного коефіцієнта детермінації R^2 для кожного допоміжного регресійного рівняння. Тому F-тест має й іншу назву – побудова допоміжної регресії. Коефіцієнт детермінації $R^2_{x_i x_1 x_2 \dots x_p}$ є коефіцієнтом детермінації в регресії, яка пов'язує фактор x_i з усіма іншими факторами. Для кожного такого коефіцієнта детермінації розраховуємо:

$$F_i = \frac{(R^2_{x_i x_1 x_2 \dots x_p})/(p-1)}{(1-R^2_{x_i x_1 x_2 \dots x_p})/(n-p)},$$

де n – кількість спостережень; p – кількість факторів.

F-тест перевіряє гіпотезу: $H_0: R^2 = 0$ проти гіпотези $H_1: R^2 \neq 0$. Розраховані значення F_i порівнюємо з критичними значеннями $F_{кр}$, за таблицями F – розподілу Фішера з *(p-1)* та *(n-p)* ступенями вільності і заданим рівнем значущості. Якщо $F_i > F_{kr}$, тоді ми відкидаємо нуль-гіпотезу і вважаємо, що фактор x_i є мультиколінеарним; якщо навпаки – то не є.

4. Характеристичні значення та умовний індекс. У деяких сучасних статистичних пакетах для перевірки наявності мульти-колінеарності використовують характеристичні значення та умовний індекс. Ми не будемо детально розглядати, як обчислювати характе-ристичні значення, бо це потребує використання апарату теорії матриць. Відзначимо лише, що за цим тестом ми розраховуємо умовне число *k*:

$$k = \frac{максимальне\ характеристичне\ значення}{мінімальне\ характеристичне\ значення}$$

і умовний індекс (CI):

$$CI = \sqrt{k}$$

Якщо $100 \le k \le 1000$, то це свідчить про помірну мульти-колінеарність, при $k > 1000$ маємо високу мультиколінеарність. Аналогічно, якщо $10 \le CI \le 30$, це свідчить про помірну мультиколінеарність, а $CI > 30$ – про високу.

Жоден із вказаних методів не є універсальним, тому бажано використовувати декілька методів одночасно.

Визначення рівня мультиколінеарності також має значення при побудові регресійних рівнянь, чим він більше – тим більша імовірність статистичної незначущості параметрів регресії. Існує показник, за допомогою якого можна визначити умовно рівень мультиколінеарності. Він носить назву ***дисперсійно-інфляційного фактору*** (*VIF – variance inflationary factor*). Для його розрахунку необхідно побудувати ряд регресійних рівнянь – залежностей відповідного фактору x_i від інших факторів моделі. Для кожного рівняння розрахувати R^2_i і потім застосувати ці значення в такій формулі:

$$VIF_i = \frac{1}{1-R_i^2}$$

Дослідники використовують значення $VIF_i = 10$ як критичне. Якщо $VIF_i \le 10$, то можна стверджувати про недостатність зв'язку між *i*-м фактором і всіма іншими. Якщо $VIF_i \ge 10$, то це свідчить про наявність мультиколінеарності.

Засоби вилучення мультиколінеарності

1. Використання первинної інформації інколи дозволяє уникнути проблеми мультиколінеарності. Для цього виявляється кількісна міра зв'язку між параметрами, відповідно замінюються фактори і отримується модель з кількістю факторів *(p-1)*, де *p* – попередня кількість факторів.

2. Метод зведення інформації, що передбачає об'єднання міжгалузевої та динамічної інформації. Цей метод був запропонований Джеймсом Тобіним (нобелівський лауреат 1981 р.).

3. Вилучення змінної (змінних) і помилка специфікації. Якщо ми маємо мультиколінеарність, ми просто можемо вилучити одну з незалежних змінних. Але вилучення змінної з моделі може призвести до помилки специфікації, що виникає через некоректне визначення моделі, що використовується в аналізі. Тоді оцінки будуть зміщені і не будуть BLUE-оцінками.

4. Перетворення змінних. Полягає в тому, що модель будується не з самих значень факторів регресії в період *t*, а з різниць їх значень (y_t-y_{t-1}), в результаті чого отримується рівняння такого виду:

$$y_t - y_{t-1} = \beta_1(x_{1t} - x_{1,t-1}) + \beta_2(x_{2t} - x_{2,t-1}) + \upsilon_t .$$

Таке рівняння отримало назву ***рівняння перших різниць.*** Цей прийом часто зменшує мультиколінеарність, бо, хоча значення x_1 та x_2 можуть мати високу кореляцію, їхні різниці не завжди висококорельовані.

Правда, такі перетворення породжують додаткові проблеми. Випадкова величина υ_t може не задовольняти припущенням моделі класичної лінійної регресії про незалежність. Ця величина іноді виявляється послідовно корельованою.

5. Збільшення спостережень може пом'якшити мультиколінеарність, якщо вибірка була невелика.

6. Розв'язанню проблеми усунення мультиколінеарності факторів може допомогти і перехід до рівняння приведеної формули. З цією метою в рівняння регресії підставляють фактор, що розглядається, виражений з іншого рівняння.

Приклад. Нехай розглядається двофакторна регресія виду: $\hat{y}_x = a + b_1x_1 + b_2x_2$, для якої фактори x_1 та x_2 мають високу кореляцію. Якщо виключити один з факторів, то ми прийдемо до рівняння парної регресії. Разом із тим можна залишити фактори в моделі, але досліджувати дане двофакторне рівняння регресії разом з іншим рівнянням, у якому фактор (наприклад, x_2), розглядається як залежна змінна.

Припустимо, що $\hat{x}_2 = A + B \cdot y + C \cdot x_3$. Підставивши це рівняння в шукане замість x_2, отримаємо:

$$\hat{y}_x = a + b_1 \cdot x_1 + b_2 \cdot (A + B \cdot y + C \cdot x_3)$$

або

$$\hat{y}_x \cdot (1 - b_2 \cdot B) = (a + b_2 \cdot A) + b_1 \cdot x_1 + C \cdot b_2 \cdot x_3$$

Якщо $(1 - b_2 \cdot B) \neq 0$, то, розділивши обидві частини рівності на $(1 - b_2 \cdot B)$, отримаємо рівняння виду:

$$\hat{y}_x = \frac{(a + b_2 \cdot A)}{(1 - b_2 \cdot B)} + \frac{b_1}{(1 - b_2 \cdot B)} \cdot x_1 + \frac{C \cdot b_2}{(1 - b_2 \cdot B)} \cdot x_3 ,$$

яке прийнято називати приведеною формою рівняння для визначення результативної ознаки *y*. Це рівняння може бути представлено у вигляді:

$$\hat{y}_x = a' + b'_1 \cdot x_1 + b'_3 \cdot x_3 .$$

До нього для оцінки параметрів може бути застосовано метод найменших квадратів.

7. Факторний аналіз.

8. Метод головних компонент.

9. Гребенева регресія.

ПИТАННЯ ДЛЯ САМОКОНТРОЛЮ

1. У чому сутність явища мультиколінеарності? Яка його природа?
2. Розкрийте, які теоретичні наслідки має мультиколінеарність.
3. Визначте практичні наслідки мультиколінеарності. Коли мультиколінеарність не є проблемою?
4. Назвіть методи виявлення мультиколінеарності, обґрунтуйте їх достатність або недостатність.
5. Які засоби вилучення мультиколінеарності ви знаєте? Чи є вони універсальними?

ПРАКТИЧНІ ЗАВДАННЯ

1. За наведеними даними перевірте наявність мультиколінеарності першим, другим та третім способами. Порівняйте результати. Зробіть висновки.

Вихідні дані до задачі 1

№	*y*	x_1	x_2
1	21	2,3	4
2	18	4,8	5
3	33	7,9	6
4	42	10,1	7
5	39	9,8	8
6	24	6,3	6
7	19	3,1	3

1. Використовуючи вихідні дані завдання № 1 вважатимемо, що фактор x_2 лінійно залежить від деякого фактору x_3, значення якого представлені нижче. Вилучіть мультиколінеарність за допомогою методу приведеної формули.

Вихідні дані до задачі 2

№	1	2	3	4	5	6	7
x_3	4,2	5,8	5,2	5,1	6,6	6,3	3,1

3. У таблиці наведено дані про виробіток продукції на добу на 10 підприємствах галузі, на який, як передбачається, впливають два фактори – витрати сировини і праці. Перевірте дані на мульти-колінеарність. При її наявності побудуйте рівняння перших різниць та перевірте отримані різниці на мультиколінеарність. Інтерпретуйте значення параметрів.

Вихідні дані до задачі 3

y	181	245	267	192	193	200	213	224	250	198
x_1	102	135	129	108	110	117	126	129	140	110
x_2	15	17	16	14	14	16	17	17	19	17

де y – обсяг виробітку продукції на добу, од.;

x_1 – витрати сировини на добу, кг;

x_2 – витрати праці, людино-діб.

ТЕСТИ

1. Ефективність оцінки параметра регресії, отриманої по МНК, означає:
   а) що вона характеризується найменшою дисперсією;
   б) що математичне сподівання залишків дорівнює нулеві;
   в) збільшення її точності із збільшенням обсягу вибірки.
2. Спроможність оцінки параметра регресії, отриманої по МНК, означає:
   а) що вона характеризується найменшою дисперсією;
   б) що математичне сподівання залишків дорівнює нулеві;
   в) збільшення її точності із збільшенням обсягу вибірки.
3. Причинами виникнення мультиколінеарності є:
   а) неоднорідність даних, що беруться за основу побудови моделі;
   б) глобальна тенденція до одночасної зміни економічних показників;
   в) інерційність незначної кількості економічних явищ і процесів.
4. Мультиколінеарність – це:
   а) явище, коли більше, ніж два фактори зв’язані між собою лінійною залежністю;
   б) явище, коли коефіцієнт кореляції між факторами має значення, що близьке або дорівнює 1;
   в) явище, коли фактори мають вплив один на одного;
   г) усі відповіді вірні.
5. Практичними наслідками мультиколінеарності є:
   а) незміщеність оцінок;
   б) ефективність оцінок;
   в) збільшення інтервалу довіри.
6. Мультиколінеарність не є проблемою, якщо:
   а) t-статистика є незначущою;
   б) єдиною метою моделювання є знаходження прогнозного значення y;
   в) єдиною метою моделювання є виявлення суттєвості зв’язку між залежною і незалежними змінними.
7. Методами виявлення мультиколінеарності є:
   а) низьке значення R^2 та незначущість t-статистики;
   б) високе значення парних коефіцієнтів кореляції;
   в) зведення інформації.
8. Методами вилучення мультиколінеарності є:
   а) використання первинної інформації;
   б) вилучення змінної;
   в) зведення інформації;
   г) усі відповіді вірні.
9. Мультиколінеарність виникає тоді, коли:
   а) помилка не має нульового середнього значення;
   б) помилка залежить від незалежної змінної;
   в) дві помилки корелюють між собою;
   г) незалежні змінні корелюють між собою;
   д) дисперсія помилок не є постійною.
10. Якщо ви оцінюєте рівняння доходів для 100 працівників, приймаючи кількість років навчання як одну незалежну змінну і кількість місяців навчання як іншу незалежну змінну, то ви можете:
    а) отримати оцінки, які не будуть BLUE;
    б) мати неоднакові дисперсії помилок;
    в) мати мультиколінеарність;
    г) б) та в).

11. Мультиколінеарність наявна, коли:
    а) дві чи більше незалежних змінних мають високу кореляцію;
    б) дисперсія випадкових величин не постійна;
    в) теперішні та лагові значення помилок корелюють;
    г) незалежна змінна виміряна з помилкою;
    д) ми будуємо неправильну версію істинної моделі.
12. Мультиколінеарність дає нам:
    а) оцінки параметрів з відхиленнями;
    б) найкращі лінійні оцінки (BLUE);
    в) неефективні оцінки параметрів;
    г) проблеми із статистичними висновками;
    д) два залишки, які корелюють один з одним.
13. Для виправлення проблеми мультиколінеарності можна:
    а) використати перехід до логарифмів;
    б) відкинути одну чи більше незалежних змінних;
    в) використати атрибутивні змінні;
    г) використати метод зважених найменших квадратів;
    д) використати залежну змінну з лагом.