

ТЕМА 11. Мультиколінеарність та її вплив на оцінювання параметрів моделі

План

11.1. Поняття та наслідки мультиколінеарності

11.2. Методи визначення та способи усунення мультиколінеарності

11.3. Алгоритм Фаррара-Глобера

11.1. Поняття та наслідки мультиколінеарності

Однією з умов застосування методу найменших квадратів для обробки різних експериментальних даних є відсутність мультиколінеарності між незалежними змінними.

В економетричну модель з багатьма змінними включають велику кількість факторів, які впливають на результативний показник. Серед цих факторів необхідно відібрати найбільш суттєві та виключити з моделі несуттєві. При цьому, виключенню з моделі підлягають ті фактори, які при парному корелюванні між собою дають високий лінійний коефіцієнт кореляції, який по абсолютній величині перевищує 0,85 ($r > 0,85$). Наявність такого лінійного зв'язку між двома факторами називається колінеарністю, а між декількома факторами – мультиколінеарністю. Таким чином, **мультиколінеарність** означає, що в багатфакторній регресійній моделі дві або більше незалежних змінних (факторів) пов'язані між собою лінійною залежністю, тобто мають високий ступінь кореляції ($r_{x_j x_{j+s}} > 0,85$).

Термін «мультиколінеарність» вперше впровадив Р. Фріш.

Мультиколінеарність (multicollinearity) – наявність тісних лінійних зв'язків між включеними до множинної економетричної моделі екзогенними змінними.

Існує повна і часткова мультиколінеарність.

Повна (екстремальна) мультиколінеарність виникає тоді, коли всі регресори пов'язані лінійною залежністю, що не дозволяє розділити внесок окремої екзогенної змінної (X_i) в залежну змінну Y . Повна колінеарність призводить до невизначеності значень параметрів. Якщо розглянути 3-вимірний простір коефіцієнтів, то в цьому просторі вектор дійсних коефіцієнтів у даному випадку не єдиний, а є цілою прямою лінією. Будь-яка точка цієї прямої – істинний вектор коефіцієнтів.

Слід проілюструвати випадок повної мультиколінеарності на прикладі двофакторної множинної регресії.

Нехай специфікація економетричної моделі має вигляд:

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \varepsilon.$$

І нехай між регресорами існує строга лінійна залежність:

$$X_2 = \alpha_0 + \alpha_1 X_1.$$

Якщо підставити в перше рівняння другий вираз, то буде отримано:

$$Y = a_0 + a_1 X_1 + a_2 (\alpha_0 + \alpha_1 X_1) + \varepsilon,$$

або

$$Y = a_0 + a_2 \alpha_0 + (a_1 + a_2 \alpha_1) X_1 + \varepsilon.$$

Введемо відповідні позначення для параметрів моделі, отримаємо специфікацію рівняння парної регресії вигляду:

$$Y = b_0 + b_1 X_1 + \varepsilon,$$

$$\text{де } b_0 = a_0 + a_2 \alpha_0,$$

$$b_1 = a_1 + a_2 \alpha_1.$$

За оцінками параметрів b_0 і b_1 неможна однозначно визначити параметри регресії (3.1), оскільки в системі (3.3) невідомих більше, ніж початкових даних.

На практиці повна мультиколінеарність зустрічається рідко.

Часткова (недосконала, стохастична) мультиколінеарність характерна для випадків, коли частина екзогенних чинників $(X_1, X_2, \dots, X_m)$

знаходиться в кореляційному зв'язку або утворює різні лінійні комбінації вигляду $X_i = a_1 X_1 + a_2 X_2 + \dots + a_m X_m$.

Регресійне рівняння задовільно описує рух залежної змінної, якщо коефіцієнт множинної кореляції досить високий, а кореляція між незалежними факторами незначна. Мультиколінеарність незалежних змінних веде до зміщення оцінок параметрів і, отже, до неможливості коректної інтерпретації результатів.

Таким чином, наявність мультиколінеарності негативно впливає на кількісні характеристики економетричної моделі або робить її побудову взагалі неможливою. Так, якщо два колінеарні вектори змінюються в одному напрямку, то майже неможливо оцінити окремий вплив кожного з них на результативний показник, оскільки кожний з цих факторів виступає лінійною комбінацією інших факторів. Наприклад, вплив на продуктивність праці таких факторів як рівень механізації та рівень енергоозброєності праці, оскільки ці фактори є колінеарними.

Причиною мультиколінеарності є наявність трендів у динамічних рядах та використання в економетричних моделях лагових значень змінних. Наприклад, при економічному зростанні більшість базових макроекономічних показників таких як дохід, споживання, накопичення, інвестиції, зайнятість мають тенденцію до

зростання з деяким лагом. Все це свідчить, що в економіці взагалі важко уникнути певного рівня залежності між показниками, тобто колінеарності.

Причини виникнення часткової мультиколінеарності:

1. Використання малої сукупності спостережень під час побудови моделі.
2. Досліджувані факторні ознаки характеризують одну і ту ж сторону явища або процесу (наприклад, показники обсягу виробленої продукції та середньорічної вартості основних засобів одночасно включати в модель не рекомендується, оскільки обидва характеризують розмір підприємства).
3. Використання факторних ознак, сумарне значення яких є постійною величиною (наприклад, коефіцієнт придатності і коефіцієнт зносу основних засобів).
4. Факторні ознаки, що є елементами один одного (наприклад, витрати на виробництво продукції і собівартість одиниці продукції).
5. Факторні ознаки, за економічним змістом дублюють одна одну (наприклад, прибуток і рентабельність продукції).
6. Факторні ознаки можуть мати загальний часовий тренд, щодо якого вони здійснюють малі коливання.
7. Значення однієї з факторних ознак є лаговим значенням іншої, тобто між зв'язком показників існує зрушення в часі.

Наслідки часткової мультиколінеарності у регресійній моделі (рис. 1):

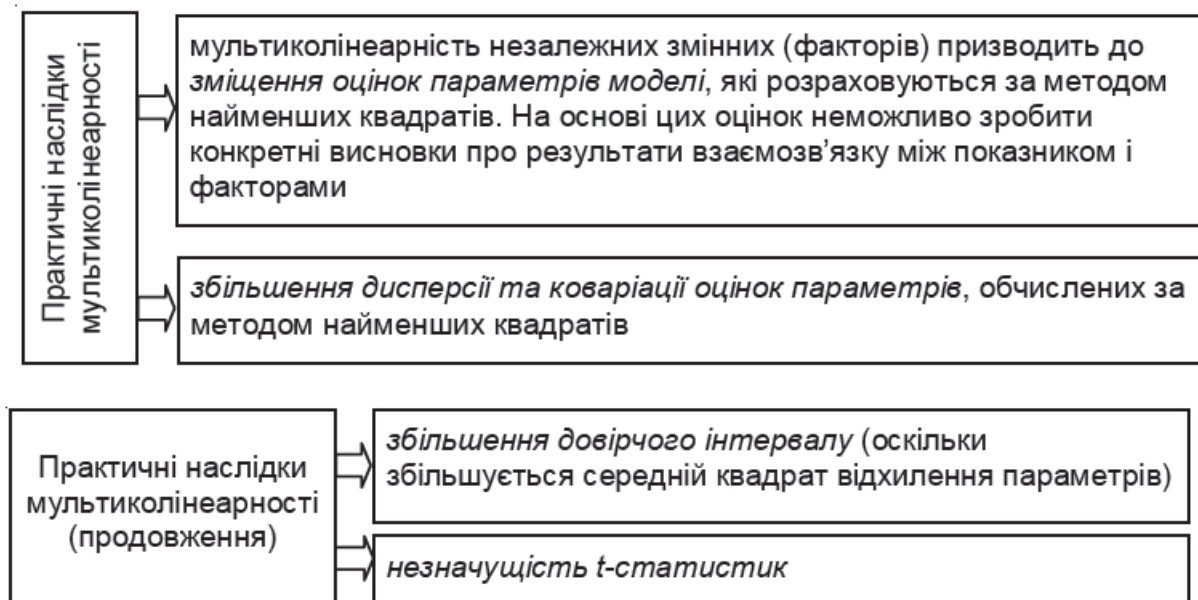


Рис. 1. Негативні наслідки присутності мультиколінеарності при розробленні економетричних моделей

1) Мультиколінеарність незалежних змінних призводить до зміщення оцінок параметрів моделі, через що з їх допомогою не можна зробити коректні висновки про результати взаємозв'язку залежної і незалежних змінних. У випадку, якщо між

незалежними змінними існує функціональний взаємозв'язок, то оцінити вплив цих змінних на залежну змінну взагалі неможливо. Тоді для оцінювання параметрів моделі метод найменших квадратів не придатний, оскільки матриця $X'X$ буде виродженою.

2) Наслідком мультиколінеарності є велика дисперсія і коваріація оцінок параметрів, обчислених за методом найменших квадратів. При наближенні коефіцієнта кореляції до свого граничного значення дисперсії параметрів зростають із значною швидкістю; так при $r=0,999$ $Var(a_1)$ у 50 разів перевищують своє значення за умови, коли немає мультиколінеарності ($r_{x1x2} = 0$).

3) Ще одним практичним наслідком мультиколінеарності є збільшення інтервалу довіри. Оскільки збільшення коефіцієнта кореляції призводить до збільшення значень середньоквадратичних відхилень параметрів, то збільшується також інтервал довіри для них. За наявності високої мультиколінеарності ($r = 0,999$) інтервал довіри в 500 разів більший, ніж коли її немає ($r = 0$).

4) Четвертим практичним наслідком мультиколінеарності є незалежність t -статистики. Щоб оцінити, чи значимо параметри багатофакторної регресії відрізняються від нуля використовують t -статистику Стюдента, тобто знаходять її

$$t_{a_j} = \frac{|a_j|}{\sigma_{a_j}}$$

фактичне значення порівнюють з його табличним значенням $t_{a,k}$. У випадку мультиколінеарності σ_{a_j} – це стандартна помилка оцінки параметру a_j

нескінченно зростає, а значення t_{a_j} прямує до нуля.

Тому при побудові багатофакторних економетричних моделей потрібно мати інформацію, що між незалежними змінними (факторами) не існує мультиколінеарності.

Мультиколінеарність не є проблемою, якщо єдиною метою регресійного аналізу є прогнозування; при цьому значення R^2 є високим, а параметри регресії – значимими, оскільки t -статистика висока. Чим вище значення R^2 , тим точніший прогноз.

Таким чином, *другий та четвертий наслідки* мультиколінеарності не дозволяють проводити аналіз економічної проблеми на основі такого регресійного рівняння, але адекватність і статистична значущість моделі в цілому дозволяють її використання для короткострокового прогнозування.

11.2. Методи визначення та способи усунення мультиколінеарності

Єдиного методу для визначення мультиколінеарності немає.

Розглянемо деякі із методів тестування мультиколінеарності. Або зовнішні ознаки мультиколінеарності (рис. 2).

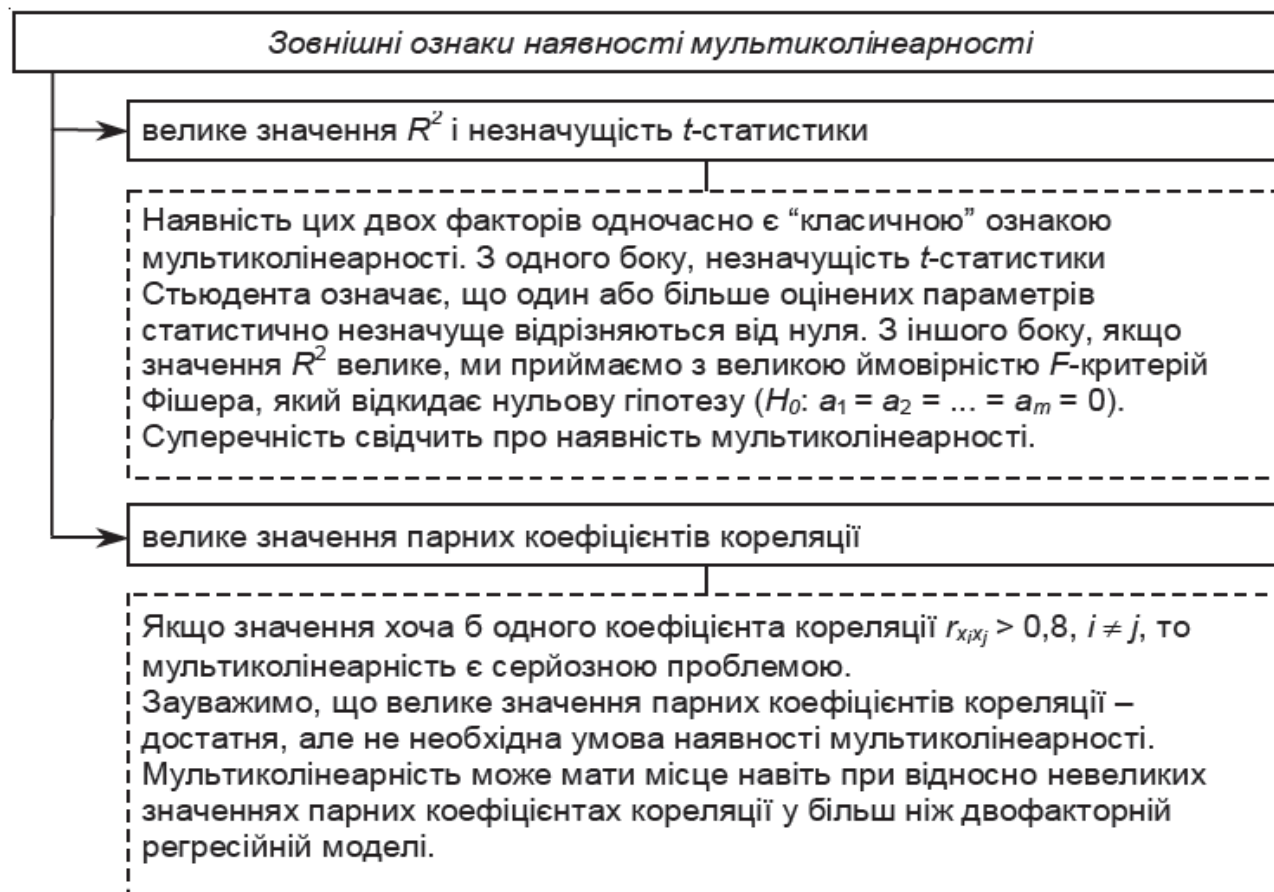


Рис. 2. Зовнішні ознаки присутності мультиколінеарності при розробці економетричних моделей

1. Високе значення R^2 і незначимість t -статистики.

Одночасна наявність цих двох факторів є «класичною» ознакою мультиколінеарності. Розглянемо m – факторну регресійну модель

$$y = a_0 + a_1 x_1 + a_2 x_2 + \dots + a_m x_m + u. \quad (11.1)$$

У випадку мультиколінеарності за t -статистикою Стюдента можна визначити, що один або більше оцінених параметрів статистично не значимо відрізняються від нуля. При високому значенні R^2 ми приймаємо з великим ступенем імовірності F -критерій Фішера, оскільки він відкидає нульову гіпотезу, коли:

$$a_1 = a_2 = \dots = a_m. \quad (5.2)$$

Тому високе значення R^2 і статистична незначимість деяких параметрів може свідчити про наявність мультиколінеарності.

2. Високе значення парних коефіцієнтів кореляції.

Другим поширеним тестом на наявність мультиколінеарності є перевірка значень парних коефіцієнтів кореляції. Якщо значення хоча б одного коефіцієнта кореляції перевищує 0,8, то в економетричній моделі має місце мультиколінеарність. Однак високе значення парних коефіцієнтів кореляції – це достатня, але не необхідна умова наявності мультиколінеарності.

Навіть при невеликих значеннях парних коефіцієнтів кореляції мультиколінеарність може бути більшою, ніж у двохфакторній регресійній моделі.

Оцінити рівень мультиколінеарності можна за допомогою величини дисперсії інфляційного VIF фактора для кожної j -ої змінної:

$$VIF_j = \frac{1}{1 - R_j^2},$$

де VIF – variance inflationary factor.

Значення $VIF_j = 10$ приймається як критичне.

Якщо $VIF_j \leq 10$, то можна стверджувати про недостатність зв'язку між j -м фактором і всіма іншими.

Якщо $VIF_j \geq 10$, то це свідчить про наявність мультиколінеарності.

Існує декілька простих способів усунення мультиколінеарності.

1. *Вилучення змінної (змінних) та помилки специфікації.* При високій мультиколінеарності найкраще та найлегше вивести з рівняння регресії одну із незалежних змінних, для якої парний коефіцієнт кореляції найбільший. Наприклад, якщо в економетричній моделі, яка описує рівень споживання від доходу та багатства, вивести змінну, що відповідає багатству, то отримаємо регресійну модель з однією незалежною змінною – доходом.

Але вилучення змінної з моделі може призвести до помилки специфікації. Помилка специфікації виникає через некоректне визначення моделі, що використовується в аналізі. Так, якщо за економічною теорією для пояснення рівня споживання модель повинна включати доход і багатство, тоді вилучення змінної багатства створюватиме помилку специфікації. Це також може призвести до зміщення оцінок.

2. *Перетворення змінних.* Однією з причин мультиколінеарності даних є їх схильність змінюватись в одному напрямку, а один із шляхів зменшення такої залежності – використання перших різниць в моделі

$$Y_t - Y_{t-1} = a_1 (x_{1t} - x_{1,t-1}) + a_2 (x_{2t} - x_{2,t-1}) + V_t,$$

$$\text{де } V = u_t - u_{t-1}.$$

Наведене вище рівняння відоме як рівняння перших різниць, оскільки ми отримали регресію не з початкових змінних, а з різниць послідовних значень змінних.

3. *Збільшення кількості спостережень.* Оскільки мультиколінеарність змінюється у кожній вибірці, то можливо в іншій моделі вона буде відсутня. Іноді просте збільшення числа спостережень у моделі, якщо це можливо, зменшує проблему мультиколінеарності.

Серед інших способів слід виділити такі: використання первинної інформації, поєднання міжгалузевої та динамічної інформації, факторний аналіз, метод головних компонент, гребенева регресія.

11.3. Алгоритм Фаррара-Глобера (Farrar – Glauber Test)

Феррар і Глобер для оцінки мультиколінеарності запропонували інші статистики за допомогою яких перевіряється мультиколінеарність не тільки всього набору незалежних факторів, але кожної із пар. Мультиколінеарність встановлюється за допомогою величини χ^2 (критерія Пірсона), розрахованої на основі оцінок взаємної кореляційної матриці незалежних змінних. Для перевірки мультиколінеарності кожної із незалежних змінних обчислюються статистики F_j .

Найповніше дослідити мультиколінеарність дає можливість алгоритм Фаррара-Глобера, який застосовує три види статистичних критеріїв для виявлення цього явища:

- 1) перевірка всього масиву незалежних змінних (критерій χ^2);
- 2) перевірка кожної незалежної змінної з усіма іншими (F-критерій);
- 3) перевірка кожної пари незалежних змінних (t-критерій) (див. рис. 3).

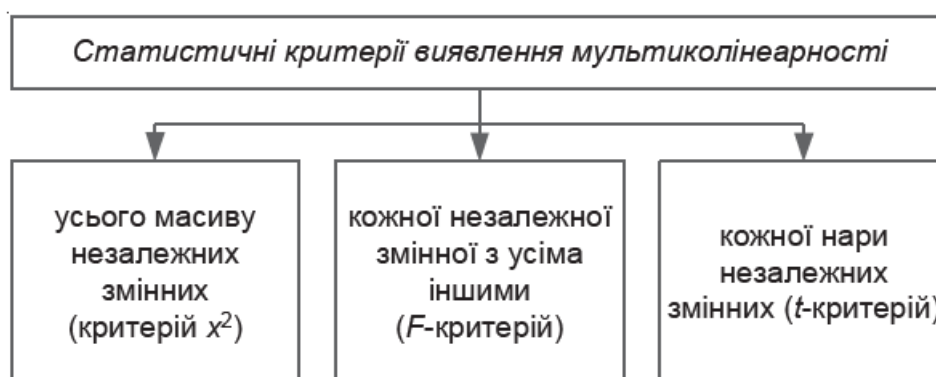


Рис. 3. Статистичні критерії ідентифікації присутності мультиколінеарності при розробленні економетичних моделей

Алгоритм метода Фаррара-Глобера

1. Нормалізація змінних.

Позначимо вектори незалежних змінних окремого регресійного рівняння через $x_1, x_2, \dots, x_j, \dots, x_m$. Елементи нормалізованих векторів обчислюємо за формулою:

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{\sqrt{n \cdot \sigma_{x_j}}} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\sum (x_i - \bar{x})^2}}, \quad i = \overline{1, n}, \quad j = \overline{1, m}, \quad (11.4)$$

де n – число спостережень для кожного вектора; m – число незалежних змінних; $\bar{x}_j$ – середня арифметична вектора $\vec{X}_j$; σ_{x_j} – стандартна похибка змінної x_j ;

$$\sigma_{x_j} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}. \quad (11.5)$$

Позначимо матрицю, складену з нормалізованих векторів через X^* .

2. Розрахунок кореляційної матриці. Взаємна кореляційна матриця оцінюється із співвідношення:

$$R = (X^{*'} \cdot X^*), \quad (11.6)$$

3. Розрахунок визначника кореляційної матриці та значення χ^2 . Якщо позначити визначник кореляційної матриці через $D = |R| = |X^{*'} \cdot X^*|$, то:

$$\chi^2 = - \left[n - 1 - \frac{1}{6}(2m + 5) \right] \cdot \ln |X^{*'} \cdot X^*|. \quad (11.7)$$

Ця величина порівнюється із табличним значенням χ^2 – квадрат критерія Пірсона для $\frac{1}{2}m(m-1)$ ступенів вільності та рівнем значимості α . Якщо $\chi^2_{\text{факт}} > \chi^2_{\text{табл}}$, то можна зробити висновок, що в масиві змінних існує мультиколінеарність.

4. Розрахунок матриці оберненої до $(X^{*'} \cdot X^*)$:

$$B = |X^{*'} \cdot X^*|^{-1} \quad (11.8)$$

5. Розрахунок F_j – статистик. Величина F_j розраховується за формулою:

$$F_j = (b_{jj} - 1) \cdot \frac{n-m}{m-1}, \quad (11.9)$$

де b_{jj} – діагональні елементи матриці B . Значення F_j порівнюється з критичними величинами F – критерія при $n-m$ і $m-1$ ступенями свободи. Якщо $F_j < F_{\alpha, k_1, k_2}$, то ні одна із незалежних змінних не мультиколінеарна з іншими.

Коефіцієнт детермінації для кожної змінної x_j розраховується так:

$$R_{x_j}^2 = 1 - \frac{1}{b_{jj}}. \quad (11.10)$$

6. Знаходження часткових коефіцієнтів кореляції.

Ці коефіцієнти розраховуються за допомогою позадіагональних елементів оберненої кореляції матриці B :

$$r_{ij}^* = \frac{-b_{ij}}{\sqrt{b_{ii} \cdot b_{jj}}}, \quad (11.11)$$

де i – номер рядка, j – номер стовпця, оберненої кореляційної матриці B .

7. Перевіряємо статистичну значущість отриманих частинних коефіцієнтів кореляції r_{ij} за допомогою t -критерію Стюдента.

Розрахунок t – статистики:

$$t_{ij} = \frac{r_{ij} \cdot \sqrt{n-m}}{\sqrt{1-r_{ij}^2}} \quad (11.12)$$

Наведені розрахунки можна інтерпретувати таким чином:

а) значення елементів матриці $(X^{*'} \cdot X^*)$ дають нам першу інформацію про мультиколінеарність незалежних факторів. Недіагональні елементи матриці дорівнюють коефіцієнтам кореляції для кожної із пар незалежних факторів і вони повинні бути меншими коефіцієнта множинної кореляції для всього рівняння;

б) мірою загальної мультиколінеарності є величина χ^2 . Цей показник не повинен перевищувати табличне значення χ_α^2 при заданому рівні значимості α і $k = \frac{1}{2}m(m-1)$ ступенях свободи;

в) величина F_j показує, які змінні мультиколінеарні. Якщо деякі із F_j більші табличних значень F – статистики при $n-m$ і $m-1$ ступенях свободи, то відповідні їм незалежні фактори мультиколінеарні;

г) за допомогою значень t_{ij} можна уточнити, які з пар незалежних факторів мультиколінеарні. Якщо деякі t_{ij} перевищують табличне значення критерія Стюдента при $n-m$ ступенях свободи, то мультиколінеарність спостерігається між парою змінних x_j і x_{j+1} .