ТЕМА 3. ОСОБЫЕ СЛУЧАИ В МНОГОФАКТОРНОМ РЕГРЕССИОННОМ АНАЛИЗЕ

Мультиколлинеарность

Количественная оценка параметров уравнения регрессии предполагает выполнение условия линейной независимости между независимыми переменными. Однако на практике объясняющие переменные часто имеют высокую степень взаимосвязи между собой, что является нарушением указанного условия. Данное явление носит название *мультиколлинеарности.*

Термин ***коллинеарность*** (*collinear*) обозначает линейную корреляцию между двумя независимыми переменными, а ***Мультиколлинеарность*** (*multi-collinear*) – между более чем двумя независимыми переменными. Обыкновенно под мультиколлинеарностью понимают оба случая.

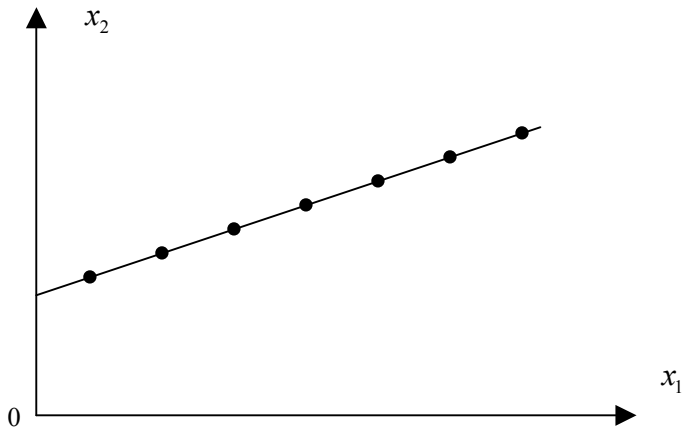
Таким образом, **мультиколлинеарность** *означает наличие тесной линейной зависимости или сильной корреляции между двумя или более объясняющими (независимыми) переменными.*

Одной из задач эконометрии является выявление мультиколлинеарности между независимыми переменными.

Различают *совершенную* и *несовершенную* мультиколлинеарность. Совершенная мультиколлинеарность означает, что вариация одной из независимых переменных может быть полностью объяснена изменением другой (других) переменной. Иначе, взаимосвязь между ними выражается линейной функцией

$$x_{i1} = \alpha_0 + \alpha_1 x_{i2},\ i = \overline{1,n}.$$

Графическая интерпретация данного случая:



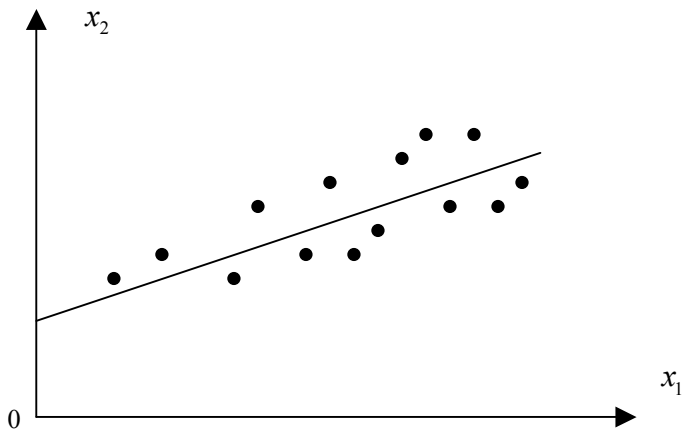
Несовершенная мультиколлинеарность может быть определена как линейная функциональная связь между двумя или более независимыми переменными, которая настолько сильна, что может существенно затронуть оценки коэффициентов при переменных в модели.

Несовершенная мультиколлинеарность возникает тогда, когда две (или более) независимые переменные находятся между собой в линейной функциональной зависимости, описываемой уравнением

$$x_{i1} = \alpha_0 + \alpha_1 x_{i2} + \varepsilon_i,\ i = \overline{1,n}.$$

В отличие от ранее рассмотренного уравнения, данное включает величину стохастической ошибки ε_i. Это предполагает, что несмотря на то, что взаимосвязь между x_1 и x_2 может быть весьма сильной, она не настолько сильна, чтобы полностью объяснить изменение переменной x_1 изменением x_2, т.е. существует некоторая необъяснимая вариация.

Графически данный случай представлен следующим образом:



В каких же случаях может возникнуть мультиколлинеарность? Их, по крайней мере, два.

1. Имеет место глобальная тенденция одновременного изменения экономических показателей. В качестве примера можно привести такие показатели как объем производства, доход, потребление, накопление, занятость, инвестиции и т.п., значения которых возрастают в период экономического роста и снижаются в период спада.

Одной из причин мультиколлинеарности является наличие тренда (тенденции) в динамике экономических показателей.

2. Использование лаговых значений переменных в экономических моделях.

В качестве примера можно рассматривать модели, в которых используются как величины дохода текущего периода, так и затраты на потребление предыдущего.

В целом при исследовании экономических процессов и явлений методами эконометрии очень трудно избежать зависимости между показателями.

Последствия мультиколлинеарности сводятся к

1. снижению точности оценивания, которая проявляется через
   a. слишком большие ошибки некоторых оценок,
   b. высокую степень корреляции между ошибками,
   c. Резкое увеличение дисперсии оценок параметров. Данное проявление мультиколлинеарности может также отразиться на получении неожиданного знака при оценках параметров;
2. незначимости оценок параметров некоторых переменных модели благодаря, в первую очередь, наличию их взаимосвязи с другими переменными, а не из-за того, что они не влияют на зависимую переменную. То есть t-статистика параметров модели не отвечает уровню значимости (t-критерий Стьюдента не выдерживает проверки на адекватность);
3. сильному повышению чувствительности оценок параметров к размерам совокупности наблюдений. То есть увеличение числа наблюдений существенно может повлиять на величины оценок параметров модели;
4. увеличению доверительных интервалов;
5. повышению чувствительности оценок к изменению спецификации модели (например, к добавлению в модель или исключению из модели переменных, даже несущественно влияющих).

Признаки мультиколлинеарности:

1. когда среди парных коэффициентов корреляции

$$r^* = \begin{pmatrix} r_{x_1/x_1} & r_{x_1/x_2} & r_{x_1/x_n} \\ r_{x_2/x_1} & r_{x_2/x_2} & r_{x_2/x_n} \\ \\ r_{x_n/x_1} & r_{x_n/x_2} & r_{x_n/x_n} \end{pmatrix}$$

между объясняющими (независимыми) переменными есть такие, уровень которых либо приближается, либо равен коэффициенту множественной корреляции.

Если в модели более двух независимых переменных, то необходимо более детальное исследование взаимосвязей между переменными. Данная процедура может быть осуществлена с помощью алгоритма Фаррара-Глобера;

2. когда определитель матрицы коэффициентов парной корреляции между независимыми переменными приближается к нулю:

   если $|r^*| = 0$, то имеет место полная мультиколлинеарность,

   если $|r^*| = 1$, то мультиколлинеарность отсутствует;

3. если в модели найдено маленькое значение параметра $\widehat{a}_j$ при высоком уровне коэффициента частной детерминации ΔR_j^2 и при этом F-критерий существенно отличается от нуля;
4. когда коэффициент частной детерминации ΔR_j^2 имеет значение, близкое к единице;
5. когда при использовании метода пошаговой регрессии вновь введенная переменная существенно изменяет оценку параметров модели при незначительном повышении значений (или их снижении) коэффициентов корреляции или детерминации;
6. когда $\overline{R}^2$ приближается к высоким значениям, близким к единице, в то время как частные значения t-критерия Стьюдента очень низки.

Мера оценки мультиколлинеарности может быть осуществлена разными способами. Один из них – расчет характеристических значений и условного индекса. Данные расчеты предлагаются некоторыми ППП по статистике. В основе вычислений лежит аппарат теории матриц.

Для того чтобы осуществить оценку уровня мультиколлинеарности рассчитывают условное число k

$$k = \frac{\text{максимальное характеристическое значение}}{\text{минимальное характеристическое значение}}$$

и условный индекс CI

$$CI = \sqrt{k}.$$

Умеренная мультиколлинеарность имеет место, если

$$100 \le k \le 1000 \text{ или}$$

$$10 \le CI \le 30.$$

Сильная – когда

$$k > 1000 \text{ или}$$

$$CI > 30 .$$

Другой способ оценки – расчет дисперсионно-инфляционного фактора VIF (VIF – *Variance Inflationary Factor*) для каждой переменной. Суть расчетов сводится к следующему: для каждой независимой переменной x_j, включенной в уравнение регрессии

$$y = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_m x_m + u ,$$

рассчитываются уравнения регрессии для независимых переменных

$$x_j = c_0 + c_1 x_1 + c_2 x_2 + \ldots + c_{j-1} x_{j-1} + c_{j+1} x_{j+1} + \ldots + c_m x_m, \ j = \overline{1,m}$$

и коэффициенты детерминации

$$R_j^2, \ j = \overline{1,m}.$$

Затем находятся дисперсионно-инфляционные факторы для каждой переменной

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

и сравниваются с критическим значением $VIF_{кр.} = 10$ (иногда $VIF_{кр.} = 5$).

Если $VIF_j \le 10$, то делают вывод о недостаточно сильной связи между j-м и остальными факторами, если $VIF_j > 10$, то делают вывод о наличии мультиколлинеарности.

Недостаток оценок мультиколлинеарности – они не дают различий между случаями, когда мультиколлинеарность существенная и когда ею можно пренебречь.

Алгоритм Фаррара-Глобера

С помощью данного алгоритма последовательно проверяется наличие мультиколлинеарности всего массива независимых переменных, каждой независимой переменной с остальными, а также попарная мультиколлинеарность.

В первом случае используется критерий χ^2 («хи»-квадрат), во втором – F-критерий Фишера и в третьем – t-критерий Стьюдента. Алгоритм распадается на семь шагов.

1-й шаг. Стандартизация (нормализация) данных.

Для каждого наблюдения всех независимых переменных осуществляются расчеты

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{\sigma_{x_j}}.$$

В результате получают векторы нормализованных данных X_j^*, которые образуют матрицу X^*.

2-й шаг. Нахождение корреляционной матрицы для независимых переменных.

Вычисляют

$$r_{x_l/x_j} = \frac{1}{n}\left(X_l^*\right)^{\prime} X_j^*$$

или в матричном виде

$$r^* = \frac{1}{n}\left(X^*\right)^{\prime} X^*,$$

где r^* – матрица коэффициентов парной корреляции независимых переменных.

3-й шаг. Вычисление значения критерия χ^2 для проверки гипотезы о наличии мультиколлинеарности всего массива данных.

Расчетное значение критерия χ^2 получается из формулы

$$\chi^2_{расч.} = -\left[n - 1 - \frac{1}{6}(2m+5)\right]\ln\left|r^*\right|,$$

где $\left|r^*\right|$ – определитель корреляционной матрицы r^*.

Данное значение χ^2-критерия сравнивается с табличным $\chi^2_{табл.}$ при числе степеней свободы $\nu = \frac{1}{2}m(m-1)$ и уровне значимости α, где m – количество независимых переменных.

Если

$$\chi^2_{расч.} > \chi^2_{табл.},$$

то в массиве данных имеет место мультиколлинеарность.

Следующие два шага позволяют исследовать наличие мультиколлинеарности между каждой независимой переменной и остальными независимыми переменными.

4-й шаг. Нахождение обратной матрицы

$$C = \left(r^*\right)^{-1}.$$

5-й шаг. Вычисление значений F-критерия Фишера для проверки гипотезы о наличии мультиколлинеарности между каждой независимой переменной и остальными независимыми переменными.

Для этого используется формула

$$F_{j\text{расч.}} = \left(c_{jj} - 1\right)\frac{n-m}{m-1},$$

где c_{jj} – диагональный элемент матрицы C.

Расчетные значения F-критерия сравниваются с табличными для числа степеней свободы $\nu_1 = n - m$ и $\nu_2 = m - 1$, и уровня значимости α. Если

$$F_{j\text{расч.}} > F_{\text{табл.}},$$

то j-я переменная мультиколлинеарна с остальными.

Для каждой переменной можно рассчитать коэффициент детерминации

$$R^2_{x_j} = 1 - \frac{1}{c_{jj}}.$$

Для оценки наличия парной мультиколлинеарности производятся действия, описанные следующими двумя шагами.

6-й шаг. Расчет частных коэффициентов корреляции.

$$r_{lj} = \frac{-c_{lj}}{\sqrt{c_{ll} \cdot c_{jj}}}.$$

Частный коэффициент корреляции *показывает тесноту связи между двумя переменными при условии, что остальные переменные постоянны, т.е. не меняются.*

7-й шаг. Расчет значений t-критерия Стьюдента для каждой пары независимых переменных.

Используется формула

$$t_{lj\text{расч.}} = \frac{r_{lj}\sqrt{n-m}}{\sqrt{1-r_{lj}^2}}.$$

Расчетные значения t-критерия сравниваются с табличным знаением при $\nu = n - m$ степенях свободы и уровне значимости α.

Если

$$t_{lj} > t_{табл.},$$

то между независимыми переменными x_l и x_j существует мультиколлинеарность.

Способы избавления от мультиколлинеарности

Для борьбы с мультиколлинеарностью можно использовать следующие способы:

1. Ничего не делать;
2. Увеличить число наблюдений;
3. Исключить из модели переменную (переменные), имеющую высокую тесноту связи с другими независимыми переменными;
4. Преобразовать мультиколлинеарные переменные путем
   - представления их в виде линейной комбинации;
   - преобразования уравнения к виду логарифмического или к уравнению в первых разностях;

   Первый прием предполагает создание новой переменной, которая является функцией мультиколлинеарных переменных и использование данной новой переменной взамен мультиколлинеарных в уравнении регрессии.

   Второй – представление мультиколлинеарной переменной в виде разности: $\Delta x_t = x_t - x_{t-1}$;
5. Использовать статистические методы: главных компонент, гребневой регрессии, факторного анализа.

Рассмотрим **пример**.

Пусть имеется выборка показателей социального развития Украины по регионам за 1997 г. Необходимо построить уравнение многофакторной регрессии для уровня безработицы и оценить наличие мультиколлинеарности массива информации. В качестве независимых переменных использовать население, номинальную среднюю зарплату и задолженность по заработной плате.

Показатели социального развития в регионах (1997 г.)

Регион	Безработица, %	Население, тыс. чел.	Номинальная средняя зарплата, грн.	Задолженность по заработной плате, млн. грн.
Автономная республика Крым	1,82	2580,2	147,6	152,3
Винницкая	2,19	1862,2	128,9	128,5
Волынская	4,23	1071,8	119,7	87,1
Днепропетровская	1,66	3811,2	194,9	370,8
Донецкая	1,84	5125,4	189,2	514,8
Житомирская	3,71	1467,9	129,2	106,3
Закарпатская	3,07	1288,6	109,0	34,7
Запорожская	1,75	2059,2	179,3	169,4
Ивано-Франковская	4,78	1465,6	129,4	89,0
Киевская	3,35	1880,4	173,5	218,2
Кировоградская	3,24	1211,2	135,6	96,7
Луганская	2,12	2743,0	160,9	272,1
Львовская	3,95	2750,6	135,8	165,5
Николаевская	2,28	1332,1	147,5	96,3
Одесская	0,47	2566,8	157,2	137,1
Полтавская	2,83	1723,9	163,9	108,0
Ровенская	3,03	1192,5	132,8	98,0
Сумская	3,08	1384,3	143,9	114,2
Тернопольская	3,49	1172,3	118,4	93,9
Харьковская	2,00	3055,1	161,1	261,6
Херсонская	1,49	1255,1	131,6	83,0
Хмельницкая	2,69	1498,4	130,3	104,0
Черкасская	2,29	1491,1	140,6	96,1
Черновицкая	1,92	940,8	123,6	53,8
Черниговская	4,22	1333,4	138,6	87,3

Использование пакета анализа позволило получить следующие результаты:

Регрессионная статистика	
Множественный R	0,565790172
R-квадрат	0,320118519
Нормированный R-квадрат	0,222992593
Стандартная ошибка	0,904684628
Наблюдения	25

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	3	8,092660189	2,697553396	3,295912152	0,040464175
Остаток	21	17,18753981	0,818454277		
Итого	24	25,2802			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	6,922226121	1,763443916	3,925401913	0,00077641	3,25494310	10,58950914
Переменная X 1	-0,000960746	0,000568251	-1,69070623	0,10568315	-0,0021424	0,000220997
Переменная X 2	-0,026075346	0,014143224	-1,84366357	0,07939056	-0,0554877	0,003337103
Переменная X 3	0,009433989	0,005805021	1,62514291	0,11904703	-0,0026382	0,021506194

Уравнение множественной регрессии выглядит следующим образом

$$y = 6{,}922 - 0{,}00096x_1 - 0{,}026x_2 + 0{,}00943x_3 .$$

Коэффициенты парной корреляции представлены в таблице

	y	x_1	x_2	x_3
y	1	-0,4279	– 0,4732	– 0,3497
x_1	– 0,4279	1	0,7631	0,9439
x_2	– 0,4732	0,7631	1	0,8133
x_3	– 0,3497	0,9439	0,8133	1

Выделим из данной матрицы только те коэффициенты, которые касаются независимых переменных. В результате получим матрицу

$$r^* = \begin{pmatrix} 1 & 0,763 & 0,944 \\ 0,763 & 1 & 0,813 \\ 0,944 & 0,813 & 1 \end{pmatrix}$$

Как видим значения всех коэффициентов больше 0,7, что свидетельствует о высокой тесноте связи между рассматриваемыми показателями. Сравнение их со значением коэффициента множественной корреляции позволяет сделать заключение о наличии мультиколлинеарности массива исходных данных.

Найдем определитель матрицы r^*

$$\left|r^*\right| = 0,03685 .$$

Значение определителя приближается к нулю, что также подтверждает ранее сделанный вывод.

Произведем оценку мультиколлинеарности путем расчета дисперсионно-инфляционных факторов. Для начала построим уравнения множественной регрессии для каждого из независимых показателей, а затем найдем множественные коэффициенты детерминации:

$$x_1 = 704,6292 - 0,6053x_2 + 8,7838x_3 ,$$

$$R_1^2 = 0,881 ,$$

$$x_2 = 120,0212 - 0,00098x_1 + 0,17897x_3 ,$$

$$R_2^2 = 0,662 ,$$

$$x_3 = -166,883 + 0,08417x_1 + 1,062379x_2 ,$$

$$R_3^2 = 0,904 .$$

Определим VIF_j :

$$VIF_1 = \frac{1}{1-0{,}881} = 8{,}4\,,$$

$$VIF_2 = \frac{1}{1-0{,}662} = 2{,}96\,,$$

$$VIF_3 = \frac{1}{1-0{,}904} = 10{,}42\,.$$

В двух случаях (первом и втором) можно сделать вывод об отсутствии мультиколлинеарности, вместе с тем, для третьей независимой переменной такого заключения сделать нельзя.

Применим алгоритм Фаррара-Глобера для более детальной оценки мультиколлинеарности.

Найдем расчетное значения критерия χ^2:

$$\chi^2_{расч.} = -\left[n-1-\frac{1}{6}(2m+5)\right]\ln\left|r^*\right| = -\left[25-1-\frac{1}{6}\cdot(2\cdot 3+5)\right]\cdot(-3{,}3) = 73{,}17\,.$$

Сравним полученное значение с табличным для числа степеней свободы $\nu = \frac{1}{2}\cdot 3\cdot(3-1) = 3$ и уровня значимости $\alpha = 0{,}05$ $\left(\chi^2_{табл.} = 7{,}81\right)$

$$73{,}17 > 7{,}81$$

и сделаем вывод о наличии мультиколлинеарности массива данных.

Определим теперь, существует ли мультиколлинеарность каждой в отдельности независимой переменной с остальными. Найдем обратную матрицу C, элементы которой используем для расчета значений F-критерия

$$C = \begin{pmatrix} 9{,}185 & 0{,}129 & -8{,}775 \\ 0{,}129 & 2{,}956 & -2{,}526 \\ -8{,}775 & -2{,}526 & 11{,}338 \end{pmatrix},$$

$$F_1 = (9{,}185-1)\cdot\frac{25-3}{3-1} = 90{,}03\,,$$

$$F_2 = (2{,}956-1)\cdot\frac{25-3}{3-1} = 21{,}52\,,$$

$$F_3 = (11{,}338-1)\cdot\frac{25-3}{3-1} = 113{,}68\,.$$

Сравним полученные значения F-критерия с табличным для $\nu_1 = n-m = 25-3 = 22$ и $\nu_2 = m-1 = 3-1 = 2$ степеней свободы, и уровня значимости $\alpha = 0{,}05$ $\left(F_{табл.} = 3{,}44\right)$

$$F_1 = 90{,}03 > F_{табл.} = 3{,}44\,,$$

$$F_2 = 21{,}52 > F_{табл.} = 3{,}44\,,$$

$$F_3 = 113{,}68 > F_{табл.} = 3{,}44$$

и, также как и в предыдущем случае, сделаем вывод о наличии мультиколлинеарности между каждой независимой переменной и остальными.

Оценим наличие парной мультиколлинеарности. Рассчитаем частные коэффициенты корреляции

$$r_{12} = \frac{-0{,}129}{\sqrt{9{,}185 \cdot 2{,}956}} = -0{,}025\,,$$

$$r_{13} = \frac{-(-8{,}775)}{\sqrt{9{,}185 \cdot 11{,}338}} = 0{,}86\,,$$

$$r_{23} = \frac{-(-2{,}526)}{\sqrt{2{,}956 \cdot 11{,}338}} = 0{,}436\,.$$

Используя полученные значения коэффициентов частной корреляции, рассчитаем t-статистики Стьюдента для каждой независимой переменной

$$t_{12\,расч.} = \frac{-0{,}025 \cdot \sqrt{25-3}}{\sqrt{1-(-0{,}025)^2}} = -0{,}116\,,$$

$$t_{13\,расч.} = \frac{0{,}86 \cdot \sqrt{25-3}}{\sqrt{1-(0{,}86)^2}} = 7{,}9\,,$$

$$t_{23\,расч.} = \frac{0{,}436 \cdot \sqrt{25-3}}{\sqrt{1-(0{,}436)^2}} = 2{,}2747\,.$$

Сравним расчетные значения t-статистик с табличным для числа степеней свободы $\nu = n - m = 25 - 3 = 22$ и уровня значимости $\alpha = 0{,}05$ $\left(t_{табл.} = 1{,}717\right)$

$$t_{12} = |-0{,}114| < 1{,}717\,,$$

$$t_{13} = 7{,}899 > 1{,}717\,,$$

$$t_{23} = 2{,}2747 > 1{,}717\,.$$

Таким образом, можно сделать окончательный вывод о том, что попарная мультиколлинеарность наблюдается между первой и третьей и между второй и третьей независимыми переменными.