

Тема 3

Особые случаи в многофакторном регрессионном анализе

- **Мультиколлинеарность**

Мультиколлинеарность

(multicollinearity) – *наличие тесной линейной зависимости или сильной корреляции между двумя или более объясняющими (независимыми) переменными*

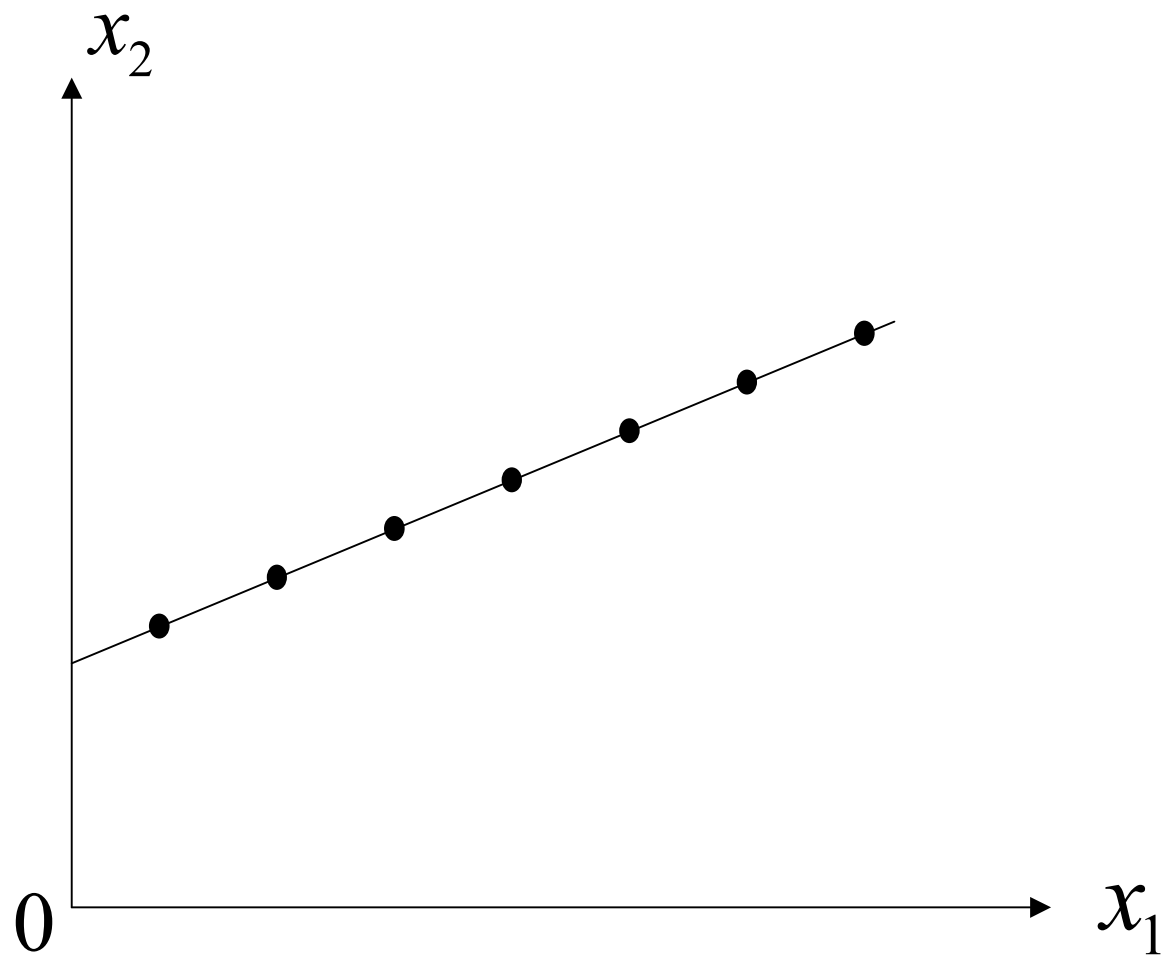
Разновидности мультиколлинеарности

- Совершенная
- Несовершенная

- Совершенная мультиколлинеарность — вариация одной из независимых переменных может быть полностью объяснена изменением другой (других) переменной (переменных)

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

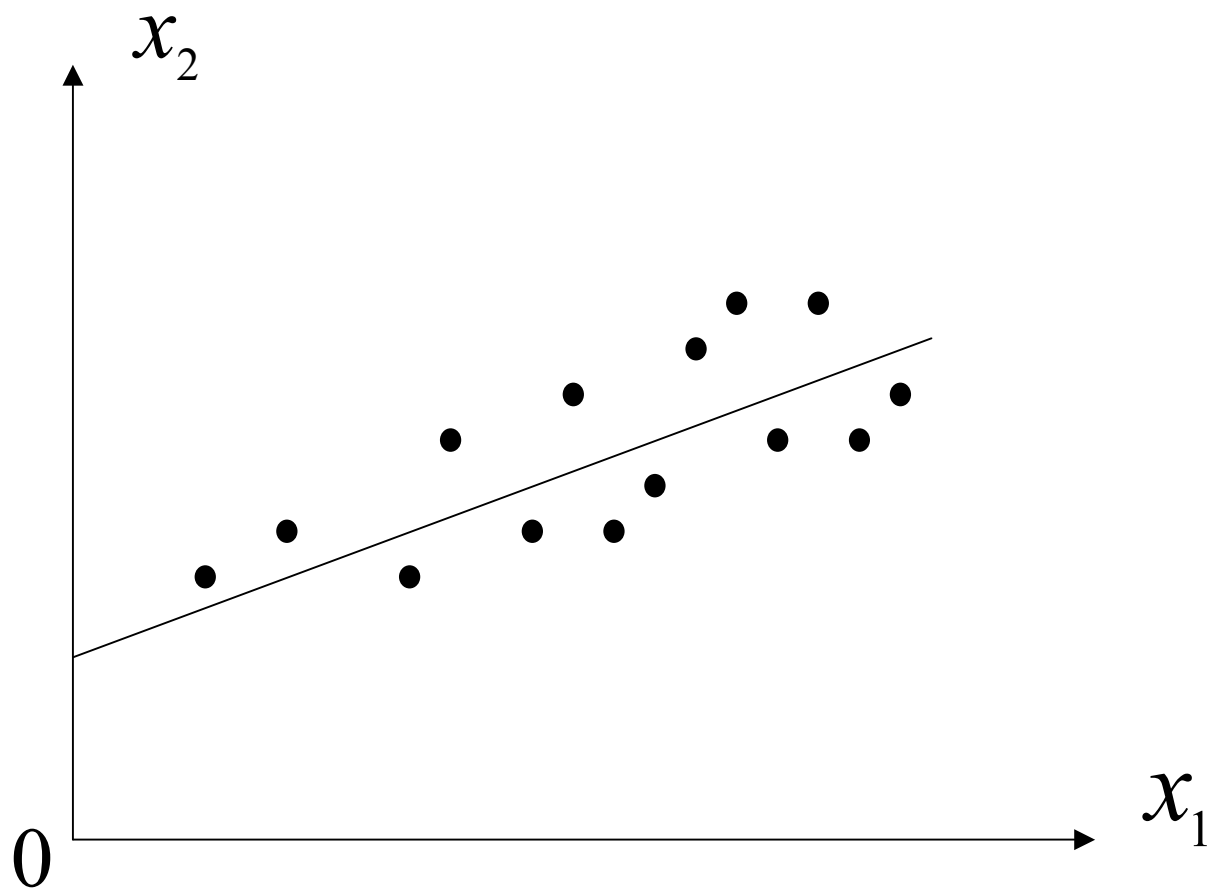
Графическая интерпретация



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

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

Графическая интерпретация



Когда возникает мультиколлинеарность?

1. Имеет место глобальная тенденция одновременного изменения экономических показателей.

Объем производства, доход, уровень потребления, уровень накопления, занятость и т.п., возрастают в период экономического роста и снижаются в период спада.

2. Использование лаговых значений переменных в экономических моделях.

Пример – модели, в которых наряду с величинами дохода текущего периода используются затраты на потребление предыдущего года.

Последствия мультиколлинеарности

- **снижение точности оценивания**
 - слишком большие ошибки некоторых оценок,
 - высокая степень корреляции между ошибками,
 - резкое увеличение дисперсии оценок параметров $\Rightarrow$ получение неожиданного знака при оценках параметров

- незначимость 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
- критерий Стьюдента не выдерживает проверки на адекватность

- сильное повышение чувствительности оценок параметров к размерам совокупности наблюдений



увеличение числа наблюдений может существенно повлиять на величины оценок параметров модели

- увеличение доверительных интервалов

- повышение 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 между независимыми переменными

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

есть такие, уровень которых либо приближается, либо равен коэффициенту множественной корреляции

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 $\hat{a}_j$ 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 $\bar{R}^2$ приближается к значениям, близким к единице, в то время как частные значения критерия Стьюдента очень низки

Методы оценки меры мультиколлинеарности

- Расчет характеристических значений и условного индекса

Рассчитывают условное число k

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

и условный индекс CI

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

Умеренная мультиколлинеарность, если

$$100 \leq k \leq 1000$$

или

$$10 \leq CI \leq 30$$

Сильная, если

$$k > 1000$$

или

$$CI > 30$$

- Расчет дисперсионно-инфляционного фактора VIF (*VIF – Variance Inflationary Factor*) для каждой переменной

Для каждой независимой переменной x_j , включенной в уравнение регрессии

$$y = a_0 + a_1x_1 + a_2x_2 + \dots + a_mx_m + u$$

рассчитываются уравнения регрессии для
независимых переменных

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

и коэффициенты детерминации

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

Дисперсионно-инфляционные факторы
для каждой переменной

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

сравниваются с критическим значением

$$VIF_{кр.} = 10$$

(иногда с $VIF_{кр.} = 5$)

Если $VIF_j \leq 10$, то связь между j -м и
остальными факторами недостаточно
сильная, если $VIF_j > 10$, то имеет место
мультиколлинеарность

- **Алгоритм Фаррара-Глобера**
- 1-й шаг.

Стандартизация (нормализация) данных

Для каждого наблюдения всех независимых переменных находим

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

Векторы нормализованных данных

$$X_j^*$$

образуют матрицу

$$X^*$$

- 2-й шаг.

Нахождение корреляционной матрицы
для независимых переменных

Вычисляют

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

или в матричном виде

$$r^* = \frac{1}{n} (X^*)' X^*$$

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

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

$|r^*|$ — определитель корреляционной матрицы

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
 m — число независимых переменных

Если

$$\chi^2_{расч.} > \chi^2_{табл.},$$

ТО В МАССИВЕ ДАННЫХ ИМЕЕТ МЕСТО
МУЛЬТИКОЛЛИНЕАРНОСТЬ

- 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

$$F_{jрасч.} = (c_{jj} - 1) \frac{n - m}{m - 1}$$

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

- 6-й шаг.

Расчет частных коэффициентов
корреляции

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

Определение

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

- 7-й шаг.

Расчет значений критерия Стьюдента для каждой пары независимых переменных

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

Расчетные значения критерия Стьюдента
сравниваются с табличным значением 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

- Ничего не делать;
- Увеличить число наблюдений;
- Исключить из модели переменную (переменные), имеющую высокую тесноту связи с другими независимыми переменными;
- Преобразовать мультиколлинеарные переменные путем

- ❖ представления их в виде линейной комбинации;
- ❖ преобразования уравнения к виду логарифмического или к уравнению в первых разностях ($\Delta x_t = x_t - x_{t-1}$)
- Использовать статистические методы: главных компонент, гребневой регрессии, факторного анализа

Пример

- 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.

Регион	Безработица, %	Нас., тыс. чел.	Ср. з.пл., грн.	Задол., мл. гр.
Авт. рес. 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
Днепроп-кая	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-Фр-кая	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

Результаты расчетов

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

Дисперсионный анализ

		<i>df</i>		<i>SS</i>		<i>MS</i>		<i>F</i>		<i>Значимость F</i>	
Регрессия		3		8,0926		2,69755		3,29591		0,040464	
Остаток		21		17,187		0,81845					
Итого		24		25,2802							

		<i>Коэф.</i>		<i>Ст. ош.</i>		<i>t-стат.</i>		<i>P-знач.</i>		<i>Нижние 95%</i>		<i>Верхние 95%</i>
У-пересеч.		6,92222		1,7634		3,9254		0,001		3,2549		10,589
Пер. X 1		-0,00096		0,0006		-1,6907		0,106		-0,002		0,0002
Пер. X 2		-0,02607		0,0141		-1,8437		0,079		-0,055		0,0033
Пер. X 3		0,00943		0,0058		1,6251		0,119		-0,003		0,0215

Уравнение множественной регрессии

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

Коэффициенты парной корреляции

	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

Матрица коэффициентов парной корреляции независимых переменных

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

Определитель матрицы r^*

$$|r^*| = 0,03685$$

Оценка мультиколлинеарность путем расчета дисперсионно-инфляционных факторов

- Строим уравнения множественной регрессии для каждого из независимых показателей,
- рассчитываем множественные коэффициенты детерминации

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

$$R_1^2 = 0,881$$

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

$$R_2^2 = 0,662$$

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

$$R_3^2 = 0,904$$

• Определяем VIF_j

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

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



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

Признак наличия мультиколлинеарности

Алгоритм Фаррара-Глобера

- расчетное значение критерия χ^2

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

Табличное значение $\chi^2_{табл.} = 7,81$

для $\nu = \frac{1}{2} \cdot 3 \cdot (3 - 1) = 3$ и $\alpha = 0,05$

$$73,17 > 7,81$$

Вывод: имеет место наличие
мультиколлинеарности массива данных

- Найдем обратную матрицу C

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

•Рассчитаем F -критерий

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

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

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

Табличное значение $F_{табл.} = 3,44$

для $\nu_1 = n - m = 25 - 3 = 22$,

$$\nu_2 = m - 1 = 3 - 1 = 2$$

и $\alpha = 0,05$

$$F_1 = 90,03 > F_{табл.} = 3,44$$

$$F_2 = 21,52 > F_{табл.} = 3,44$$

$$F_3 = 113,68 > F_{табл.} = 3,44$$

Вывод: имеет место мультиколлинеарность между каждой независимой переменной и остальными

- Оценим наличие парной мультиколлинеарности

Рассчитаем частные коэффициенты корреляции

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

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

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

- Рассчитаем статистики Стьюдента для каждой независимой переменной

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

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

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

Табличное значение $t_{табл.} = 1,717$

для $\nu = n - m = 25 - 3 = 22$

и $\alpha = 0,05$

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

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

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

Вывод: попарная мультиколлинеарность наблюдается между первой и третьей и между второй и третьей независимыми переменными