Мультиколлинеарность

Лекция проф. Орловой Ирины Владленовны

Мультиколлинеарность - тесная корреляционная взаимосвязь между отбираемыми для анализа факторами, совместно воздействующими на общий результат.

Виды мультиколлинеарности. Строгая и нестрогая мультиколлинеарность

1. Строгая (полная, функциональная) мультиколлинеарность - наличие линейной функциональной связи между объясняющими переменными.

2. Нестрогая (частичная) мультиколлинеарность - наличие сильной линейной корреляционной связи между объясняющими переменными .
3. Чем ближе мультиколлинеарность к строгой (совершенной), тем серьезнее ее последствия.

Линейная зависимость двух или нескольких регрессоров называется мультиколлинеарностью

Функциональная зависимость между регрессорами называется полной мультиколлинеарностью

Стохастическая зависимость между регрессорами называется частичной мультиколлинеарностью

Предпосылка Гаусса-Маркова относительно матрицы регрессоров *X*

$$rank(X) = rank(X^T X) = k$$

k – число столбцов матрицы регрессоров

Полная мультиколлинеарность соответствует случаю, когда предположение, что матрица $(X'X)$ невырождена, т. е. ее определитель отличен от нуля: $\det X^T X \neq 0$ нарушается, т. е. когда столбцы матрицы линейно зависимы.

Это приводит к невозможности решения соответствующей системы нормальных уравнений и получения оценок параметров регрессионной модели.

1. Корреляционные связи есть всегда. Проблема мультиколлинеарности - сила проявления корреляционных связей.

2. Однозначных критериев мультиколлинеарности не существует.

3. Строгая мультиколлинеарность делает построение регрессии невозможным. (Согласно теоремы Кронекера-Капелли система уравнений имеет бесчисленное множество решений).

Мультиколлинеарность проявляется в совместном действии факторов:

1. Построить модель - значит определить вклад каждого фактора.

2. Если два или более фактора изменяются только совместно, их вклад по отдельности становится невозможно различить.

3. Чем более сильно коррелированны переменные, тем труднее различить их вклад.

В экономических исследованиях мультиколлинеарность чаще проявляется в нестрогой (стохастической) форме, когда между хотя бы двумя объясняющими переменными существует тесная корреляционная связь. Определитель матрицы X'X не равен нулю, но очень мал. В этом случае затрудняется экономическая интерпретация параметров уравнения регрессии, так как **некоторые из его коэффициентов могут иметь неправильные с точки зрения экономической теории знаки** и **неоправданно большие значения**. **Оценки параметров ненадежны**, **обнаруживают большие стандартные ошибки** и меняются с изменением объема наблюдений (не только по величине, но и по знаку), что делает модель непригодной для анализа и прогнозирования

ВНИМАНИЕ!

Рассматриваемые в презентации примеры можно найти в

«Экономико-математические методы и модели: компьютерное моделирование: Учебное пособие– 3-е изд., перераб. и доп. / И.В. Орлова, В.А. Половников. – М.: Вузовский учебник: ИНФРА-М, 2014.» / ЭБС ZNANIUM.COM

Рекомендуемая литература по теме

Обнаружение мультиколлинеарности

- Один из подходов заключается в анализе матрицы коэффициентов парной корреляции. Считают явление мультиколлинеарности в исходных данных установленным, если коэффициент парной корреляции между двумя переменными больше 0,8.

- Другой подход состоит в исследовании матрицы X'X. Если определитель матрицы X'X близок к нулю, то это свидетельствует о наличии мультиколлинеарности.

$$|\mathbf{R}| = \begin{vmatrix} 1 & r_{12} & \dots & r_{1m} \\ r_{21} & 1 & \dots & r_{2m} \\ \dots & \dots & \dots & \dots \\ r_{m1} & r_{m2} & \dots & 1 \end{vmatrix} \to 0$$

ПРИМЕР. Задача состоит в построении модели для предсказания объема реализации одного из продуктов кондитерской фирмы.
Объем реализации – это зависимая переменная Y(млн. руб.) В качестве независимых, объясняющих переменных выбраны: время - X1, расходы на рекламу X 2 (тыс. руб.), цена товара X3 (руб.), средняя цена товара у конкурентов X4 (руб.), индекс потребительских расходов X5 (%).

	Y	*X1*	*X2*	*X3*	*X4*	*X5*
Y	**1**					
X1	**0,678**	**1**				
X2	**0,646**	**0,106**	**1**			
X3	**0,233**	**0,174**	**-0,003**	**1**		
X4	**0,226**	**-0,051**	**0,204**	**0,698**	**1**	
X5	**0,816**	**0,960**	**0,273**	**0,235**	**0,031**	**1**

	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%	Верхние 95%
Y-пересечение	-3017,40	1094,49	-2,76	0,020	-5456,061	-578,731
X1	-13,42	10,38	-1,29	0,225	-36,544	9,706
X2	6,67	3,01	2,22	0,051	-0,032	13,376
X3	-6,48	15,78	-0,41	0,69	-41,63	28,68
X4	12,24	14,41	0,85	0,416	-19,868	44,345
X5	30,48	11,52	2,64	0,025	4,797	56,154

Обнаружение мультиколлинеарности

1. Высокие коэффициенты детерминации и F-статистика, но некоторые (или даже все) коэффициенты незначимы, т.е. имеют низкие t - статистики.
2. Высокие парные коэффициенты корреляции.
3. Высокие частные коэффициенты корреляции.
4. Высокие значения коэффициента VIF («фактор инфляции вариации»).
5. Знаки коэффициентов регрессии противоположны ожидаемым.
6. Добавление или удаление наблюдений из выборки сильно изменяют значения оценок.

Обнаружение мультиколлинеарности. Анализ матрицы коэффициентов парной корреляции

ПРИМЕР. Задача состоит в построении модели для предсказания объема реализации одного из продуктов кондитерской фирмы.

Объем реализации – это зависимая переменная Y(млн. руб.) В качестве независимых, объясняющих переменных выбраны: время - X1, расходы на рекламу X 2 (тыс. руб.), цена товара X3 (руб.), средняя цена товара у конкурентов X4 (руб.), индекс потребительских расходов X5 (%).

	Y	*X1*	*X2*	*X3*	*X4*	*X5*
Y	1					
X1	0,678	1				
X2	0,646	0,106	1			
X3	0,233	0,174	-0,003	1		
X4	0,226	-0,051	0,204	0,698	1	
X5	0,816	0,960	0,273	0,235	0,031	1

$$y = f(x_2, x_5)$$

$$\hat{Y} = -1471.31 + 9.56X_2 + 15.75X_5$$

Обнаружение мультиколлинеарности Тест Фаррара-Глоубера

Этот алгоритм содержит три вида статистических критериев проверки наличия мультиколлинеарности:

1) всего массива переменных (критерий «хи-квадрат»);
2) каждой переменной с другими переменными (F-критерий);
3) каждой пары переменных (t-тест).

Обнаружение мультиколлинеарности
Тест Фаррара-Глоубера (1)
Проверка наличия мультиколлинеарности всего массива переменных (критерий «хи-квадрат»)

- Построить корреляционную матрицу R и найти её определитель
- Вычислить наблюдаемое значение статистики Фаррара - Глобера по следующей формуле:

$$FG_{набл} = n - \left[-1 - \frac{1}{6}k(2 + 5) \right] \ln(\det[R]),$$

Эта статистика имеет распределение χ^2 (хи-квадрат).

- Фактическое значение этого критерия сравнивается с табличным значением χ^2 с 0,5k(k-1) степенями свободы и уровне значимости α. Если $FG_{набл}$ больше табличного, то в массиве объясняющих переменных существует мультиколлинеарность.

Обнаружение мультиколлинеарности. Тест Фаррара-Глоубера (1)

Проверка наличия мультиколлинеарности всего массива переменных

	X2	*X3*	*X4*	*X5*
X2	1	-0,003	0,204	0,273
X3	-0,003	1	0,698	0,235
X4	0,204	0,698	1	0,031
X5	0,273	0,235	0,031	1

Вычислим определитель матрицы R = 0,373. Вычислим *FGнабл:*

$$FG_{набл} = -\left[n - 1 - \frac{1}{6}(2k+5)\right]\ln(\det[R]) =$$

$$= -\left[16 - 1 - \frac{1}{6}(2\cdot 4 + 5)\right]\ln(0,373) = 12,66$$

FGнабл > FGкрит=12,59 *Но* отклоняется, факторы признаются коллинеарными

Обнаружение мультиколлинеарности. Тест Фаррара-Глоубера (2)

Проверка наличия мультиколлинеарности каждой переменной с другими переменными (F-критерий)

1. Вычислить обратную матрицу $C = R^{-1}$
2. Вычислить F-критерии

$$F_j = \left(c_{jj} - 1\right)\frac{n-k-1}{k}$$

где c_{jj} – диагональные элементы матрицы С.

3 Фактические значения F критериев сравнить с табличным значением при $v_1 = k$ и $v_2 = (n - k - 1)$ степенях свободы и уровне значимости α, где k – количество факторов. Если $F_j > Fтабл$ то соответствующая j-тая независимая переменная мультиколлинеарна с другими.

Обнаружение мультиколлинеарности. Тест Фаррара-Глоубера (2)

Проверка наличия мультиколлинеарности каждой переменной с другими переменными (F-критерий)

1.Вычислим обратную матрицу $C = R^{-1}$

	X2	X3	X4	X5
X2	1,2518	0,544	-0,621	-0,451
X3	0,544	2,37586	-1,749	-0,654
X4	-0,621	-1,749	2,331	0,510
X5	-0,451	-0,654	0,510	1,26162

2. Вычислим F-критерии

F2	F3	F4	F5
0,692	3,7836	3,6603	0,719

3. Табличное значение F-критерия = **3,36**

4. 3 и 4-ая независимые переменные мультиколлинеарнаы с другими.

Обнаружение мультиколлинеарности. Тест Фаррара-Глоубера (3)

Проверка наличия мультиколлинеарности каждой пары переменных (t-тест).

- Найти частные коэффициенты корреляции:

$$r_{ij(\ )} = \frac{-c_{ij}}{\sqrt{c_{ii} \cdot c_{jj}}}$$

- где – c_{ij} элемент матрицы C, содержащийся в i-ой строке и j-ом столбце; c_{ii} и c_{jj} – диагональные элементы матрицы C.
- Вычисление t-критериев:

$$t_{ij} = \frac{r_{ij(\ )}\sqrt{n-k-1}}{\sqrt{1-r^2_{ij(\ )}}}$$

Фактические значения критериев t_{ij} сравниваются с табличным $t_{табл}$ при $(n-k-1)$ степенях свободы и уровне значимости α.

Если $|t_{ij}| > t_{табл}$, то между независимыми переменными i и j существует мультиколлинеарность.

Обнаружение мультиколлинеарности. Тест Фаррара-Глоубера (3)

Проверка наличия мультиколлинеарности каждой пары переменных (t-тест).

	Частные коэффициенты корреляции	Вычисление t-критериев:
X2 X3	-0,315	-1,102
X2 X4	0,363	1,294
X2 X5	0,359	1,275
X3 X4	0,743	3,682
X3 X5	0,378	1,353
X4 X5	-0,297	-1,032

Табличное значение t критерия = 2,2. $|t_{3,4}| > t_{табл}$

Удаляем X3 т.к. у него больше значение F критерия. Остаются X2, X4, X5

$$\hat{Y} = -1654.76 + 9.05X_2 + 10.54X_4 + 15.82X_5$$

Фактор инфляции вариации как оценка эффекта мультиколлинеарности

Для измерения эффекта мультиколлинеарности используется показатель VIF – «фактор инфляции вариации»

$$VIF_{x_j} = \frac{1}{\left(1 - R^2_{x_j x_1 \ldots x_{j-1} x_{j+1} \ldots x_p}\right)}$$

Обнаружение мультиколлинеарности. Метод инфляционных факторов

Алгоритм метода заключается в следующем:

1.Строятся уравнения регрессии, которые связывают каждый из регрессоров со всеми оставшимися.

2. Вычисляются коэффициенты детерминации R^2 для каждого уравнения регрессии.

3. Проверяется статистическая гипотеза H0: $R^2=0$ с помощью F теста. Вывод: если гипотеза H0: $R^2=0$ не отклоняется, значит данный регрессор не приводит к мультиколлинеарности.

4. Значения VIF j > 10.0 могут указывать на наличие мультиколлинеарности.

Обнаружение мультиколлинеарности. Метод инфляционных факторов

Минимальное возможное значение = 1.0
Значения VIF j > 10.0 могут указывать на наличие мультиколлинеарности

X1	21,112
X2	1,889
X3	2,474
X4	2,331
X5	23,389

	Y	*X1*	*X2*	*X3*	*X4*	*X5*
Y	1					
X1	0,678	1				
X2	0,646	0,106	1			
X3	0,233	0,174	-0,003	1		
X4	0,226	-0,051	0,204	0,698	1	
X5	0,816	0,960	0,273	0,235	0,031	1

Обнаружение мультиколлинеарности. Метод Белсли.

Для определения мультиколлинеарности используем метод Белсли. Belsley, Kuh и Welsch предложили метод анализа мультиколлинеарности основанный на индексах обусловленности (the scaled condition indexes) и дисперсионных долях (the variance-decomposition proportions). Обусловленность оценивает близость матрицы коэффициентов к вырожденной. Число обусловленности η является количественной оценкой обусловле $\eta_j = \mu_{\max} / \mu_j$. Отметим, что всегда η> 1. Если η > 103, то говорят, что матрица плохо обусловлена. Если 1 < η < 100 , то матрица считается хорошо обусловленной.

Оценки собственных значений

$\lambda_1 \approx 0.334877595627432$

$\lambda_2 \approx 3.48909013788591$

$\lambda_3 \approx 14.7339269862456$

$\lambda_4 \approx 24.0773778440838$

$\lambda_5 \approx 32.3647274361573$

$\eta_j = \mu_{\max} / \mu_j$ **96,65=32,36/0,33**

Методы устранения мультиколлинеарности

1. Изменить или увеличить выборку.
2. **Исключить из модели одну или несколько переменных.**
3. Преобразовать мультиколлинеарные переменные: - использовать нелинейные формы; - использовать агрегаты (линейные комбинации переменных); - использовать первые разности вместо самих переменных.
4. Использовать при оценке коэффициентов метод главных компонент или другие специальные процедуры расчета коэффициентов при плохой обусловленности x'x.
5. **Использовать пошаговые процедуры отбора факторов.**