Мультиколлинеарность

Теоретическая (строгая) мультиколлинеарность

Теоретическая мультиколлинеарность данных – явление, наблюдаемое при нарушении условий теоремы Гаусса – Маркова об отсутствии точной линейной связи между регрессорами. При наличии теоретической мультиколлинеарности однозначное нахождение оценов МНК коэффициентов регрессии невозможно.

Теоретическая (строгая) мультиколлинеарность

Возникновение строгой мультиколлинеарности означает, что на начальном этапе набор независимых переменных был сформирован некорректно

Пример: включение в правую часть одного уравнения экспорта, импорта и чистого экспорта

Теоретическая мультиколлинеарность

Вспомним: оцененное регрессионное уравнение в матичном виде:

$$Y = X\hat{\beta} + e$$

Решение этого матричного уравнения:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Условие существования обратной матрицы:

$$\Delta(X^T X) \neq 0$$

Теоретическая мультиколлинеарность

Условие существования обратной матрицы:

$$\Delta(X^T X) \neq 0$$

Вопрос: **когда это условие нарушается, т. е. при каких условиях** $\Delta(X^T X) = 0$?

Ответ: если среди столбцов матрицы X есть линейно зависимые.

$$X = \begin{pmatrix} 1 & X_{11} & \dots & X_{k1} \\ 1 & X_{12} & \dots & X_{k2} \\ \dots & \dots & \dots & \dots \\ 1 & X_{1n} & \dots & X_{kn} \end{pmatrix}$$

Линейная зависимость векторов

Определение. Пусть $\boldsymbol{z_1}, \ldots, \boldsymbol{z_s} \in \mathbb{R}^n$. Если существуют действительные числа $\lambda_1, \ldots, \lambda_s$, не все равные нулю, для которых имеет место

$$\boldsymbol{z_1}\lambda_1 + \cdots + \boldsymbol{z_s}\lambda_s = 0,$$

то векторы $\boldsymbol{z_1}, \ldots, \boldsymbol{z_s}$ называют **линейно зависимыми**.

Векторы $\boldsymbol{z_1}, \ldots, \boldsymbol{z_s}$ - **линейно независимые**, если из $\boldsymbol{z_1}\lambda_1 + \cdots + \boldsymbol{z_s}\lambda_s = 0$ следует $\lambda_1 = \lambda_2 = \cdots = \lambda_s = 0$.

Пример. Векторы $\mathbf{a} = (1,\ 2,\ 3)$, $\mathbf{b} = (4,\ 5,\ 6)$, $\mathbf{c} = (7, 8, 9)$ линейно зависимы. Поскольку $1\mathbf{a} - 2\boldsymbol{b} + 1\boldsymbol{c} = \mathbf{0}$.

Векторы $\mathbf{a} = (1,\ 0,\ 0)$, $\mathbf{b} = (0,\ 1,\ 0)$, $\mathbf{c} = (0,\ 0,\ 1)$ - линейно независимы. **Проверьте!**

Теорема. *Определитель квадратной матрицы равен нулю тогда и только тогда, когда строки (столбцы) этой матрицы линейно зависимы.*

Пример. Выяснить, являются ли следующие строки линейно зависимыми:

$$A_1 = (1, 2, -4),\ A_2 = (2, -1, 0),\ A_3 = (4, 3, -8).$$

Решение.

$$\begin{vmatrix} 1 & 2 & -4 \\ 2 & -1 & 0 \\ 4 & 3 & -8 \end{vmatrix} = 0$$

Вывод. **Данные строки линейно зависимы.**

Теоретичекая мультиколлинеарность

$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_K X_K + \varepsilon$,

Теоретическая мультиколлинеарность: Rank(X) < k + 1

Ex.1.

$$\ln wage = \beta_1 + \beta_2 S + \beta_3 MALE + \beta_4 FEMALE + \ldots + \varepsilon,$$

$FEMALE_i + MALE_i = 1$ *(для каждого наблюдения)*

Ex.2.

$$\ln price = \beta_1 + \beta_2 livsq + \beta_3 nonlivsq + \beta_4 totsq + \ldots + \varepsilon,$$

$$livsq + nonlivsq = totsq$$

*Rank(X) – ранг матрицы (X). Ранг матрицы равен числу линейно независимых строк (столбцов).

Теоретическая мультиколлинеарность

Пример теоретической мультиколлинеарности

Ex.3.

$$\Pr ice = \beta_1 + \beta_2 D_I + \beta_3 D_{II} + \beta_4 D_{III} + \beta_5 D_{IV} + \ldots + \varepsilon,$$

$$D_I + D_{II} + D_{III} + D_{IV} = 1 \text{ (для каждого наблюдения)}$$

Dummy trap

Частичная (неполная) мультиколлинеарность

При работе с реальными данными часто возникает **частичная (неполная) мультиколлинеарность**, когда между регрессорами существует **ПОЧТИ линейная зависимость**.

В этом случае

$$\Delta(X^T X) \approx 0$$

***Далее под термином «мультиколлинеарность» будем понимать частичную мультиколлинеарность*

Последствия мультиколлинеарности

— Основное негативное последствие — стандартные ошибки оценок коэффициентов оказываются высокими. **Точность** оценивания **оказывается низкой**

— Оценки коэффициентов остаются **несмещенными**

— Та или иная степень корреляции между регрессорами существует всегда. Проблема возникает только когда эта линейная связь проявляется слишком сильно

Выявление м/к на начальном этапе моделирования (до регрессии)

Индикаторы мультиколлинеарности

- **В корреляционной матрице факторов встречаются элементы, по модулю близкие к 1** (по модулю больше 0,7)

- **Достаточно большое значение** (больше 6)

VIF – variance inflation factor хотя бы для одного фактора

$$VIF(X_j) = \frac{1}{1 - R_j^2},$$

где R_j^2 – коэффициент множественной детерминации регрессора X_j на все остальные регрессоры.

Пример 1 (корр. матрица)

Рассмотрим пример.

Изучается зависимость потребительских расходов от богатства и дохода.

consumption	wealth	income	ln(consump)	ln(wealth)	ln(income)
70	810	80	4,248495242	6,697034248	4,382026635
65	1009	100	4,17438727	6,91671502	4,605170186
90	1237	120	4,49980967	7,120444372	4,787491743
95	1425	140	4,553876892	7,261927093	4,941642423
110	1633	160	4,700480366	7,398174093	5,075173815
115	1876	180	4,744932128	7,53689713	5,192956851
120	2025	200	4,787491743	7,61332498	5,298317367
140	2201	220	4,941642423	7,696667082	5,393627546
155	2435	240	5,043425117	7,797702036	5,480638923
150	2686	260	5,010635294	7,895808377	5,560681631

Пример 1 (корр. матрица)

Матрица парных коэфициентов корреляции

	ln(consump)	*ln(wealth)*	*ln(income)*
ln(consump)	1		
ln(wealth)	0,97330579	1	
ln(income)	0,973641809	0,999514773	1

Принцип исключения факторов:

- Если две переменные явно коллинеарны ($r_{x_i x_j} \geq 0{,}7$), то одну из них исключаем.
- Включаем фактор, имеющий наименьшую тесноту связи с другими факторами

Расчет корреляционной матрицы в Excel

Пример (VIF)

Оцененная с помощью МНК зависимость заработной платы индивида EARN от его возраста AGE, опыта работы EXP, пола MALE, длительности обучения S имеет вид:

$EARN = -24 - 0{,}099*AGE + 2{,}49*S + 0{,}46*EXP + 6{,}23*MALE,\ R^2 = 0{,}428.$

Были также оценены вспомогательные регрессии:

$AGE = -0{,}007 + 0{,}53*S + 0{,}23*EXP + 1{,}23*MALE,\ R^2 = 0{,}88;$

$S = 8{,}47 + 0{,}095*AGE - 0{,}2*EXP + 0{,}12*MALE,\ R^2 = 0{,}26;$

$EXP = -0{,}07 + 0{,}53*AGE - 0{,}6*S + 1{,}23*MALE,\ R^2 = 0{,}93,$

Найдем VIF для переменных AGE, S, EXP.

$$VIF(AGE) = \frac{1}{1 - R^2_{AGE}} = \frac{1}{1 - 0{,}88} = 8{,}33$$

$$VIF(S) = \frac{1}{1 - R^2_{S}} = \frac{1}{1 - 0{,}26} = 1{,}35$$

$$VIF(AGE) = \frac{1}{1 - R^2_{AGE}} = \frac{1}{1 - 0{,}93} = 14{,}28.$$

Вывод: мультиколлинеарность есть

Выявление м/к в построенной модели («симптомы» м/к)

- Небольшие изменения в данных приводят к значительным изменениям в оценках коэффициентов регрессии.

- Многие коэффициенты по-отдельности не значимы, хотя в целом регрессия адекватная, R^2 может быть достаточно высоким.

- Оценки коэффициентов регрессии (обычно незначимых) могут иметь “неправильный” знак (с экономической точки зрения).

Выявление м/к в построенной модели («симптомы» м/к)

Таблица 4.1

Dependent Variable: LOG(CONSUMP)
Method: Least Squares
Date: 02/27/07 Time: 16:42
Sample: 1 10
Included observations: 10

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.608456	4.899733	0.124181	0.9047
LOG(INCOME)	0.643406	2.137046	0.301073	0.7721
LOG(WEALTH)	0.108048	2.126388	0.050813	0.9609

R-squared	0.947998	Mean dependent var	4.670518
Adjusted R-squared	0.933140	S.D. dependent var	0.300991
S.E. of regression	0.077828	Akaike info criterion	-2.025296
Sum squared resid	0.042401	Schwarz criterion	-1.934521
Log likelihood	13.12648	F-statistic	63.80453
Durbin-Watson stat	2.811065	Prob(F-statistic)	0.000032

Таблица 4.2

Dependent Variable: LOG(CONSUMP)
Method: Least Squares
Date: 02/27/07 Time: 17:23
Sample: 1 9
Included observations: 9

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.345546	5.649167	-0.061168	0.9532
LOG(INCOME)	0.276879	2.422325	0.114303	0.9127
LOG(WEALTH)	0.489127	2.423483	0.201828	0.8467

R-squared	0.940169	Mean dependent var	4.632727
Adjusted R-squared	0.920226	S.D. dependent var	0.293008
S.E. of regression	0.082758	Akaike info criterion	-1.884587
Sum squared resid	0.041093	Schwarz criterion	-1.818845
Log likelihood	11.48064	F-statistic	47.14140

Устранение мультиколлинеарности

- Увеличить число наблюдений
- Исключить переменную, с которой связана мультиколлинеарность.
  Следует помнить, что иногда это может привести к более серьезным проблемам
- Использовать нелинейные формы зависимостей
- Использовать агрегаты: линейные комбинации переменных
- А может, ничего не делать? ☺