ЛЕКЦИЯ 12

НАРУШЕНИЯ ПРЕДПОСЫЛОК ТЕОРЕМЫ ГАУССА-МАРКОВА:
Ч. I. МУЛЬТИКОЛЛИНЕАРНОСТЬ

1. Что такое мультиколлинеарность? Причины ее возникновения
2. Проявления и последствия мультиколлинеарности
3. «Измерители» мультиколлинеарности
4. Что же с ней в итоге делать?!

Вспомним основные предположения теоремы Гаусса-Маркова:

Если:

1) Модель $Y = X\beta + u$ специфицирована верно;
2) X – детерминированная матрица, причем $rank(X) = k$;
3) $E(u) = 0$ и $Var(u) = E(u'u) = \sigma^2 I_n$,
4) $E(u_i u_j) = 0, при\ \ i \neq j$,

то оценка МНК $b = \hat{\beta} = (X'X)^{-1} X'Y$ является наиболее эффективной оценкой в классе линейных несмещенных оценок (BLUE).

К чему приводят ошибки спецификации модели, мы посмотрели в предыдущей лекции.

Нарушение предположения (2) приводит к мультиколлинеарности;
нарушение (3) – к гетероскедастичности,
а нарушение (4) ведет к автокорреляции случайного возмущения.

Сегодня проанализируем влияние мультиколлинеарности на качество построенной модели.

Что такое мультиколлинеарность? Причины ее возникновения

Находя МНК-оценки параметров модели, мы предполагали, что X – детерминированная матрица полного ранга, т.е. $rank(X)=k$, где k – количество столбцов матрицы X (объясняющие переменные линейно независимы).

Что будет, если это условие нарушено?

Тогда $rank(X)<k$, следовательно, $\det(X'X)=0$, и мы не сможем обратить матрицу $(X'X)$, а значит, не сможем найти оценки параметров регрессии.

Такая ситуация называется точной (полной, совершенной) мультиколлинеарностью (perfect collinearity).

Рассмотрим на примере:

Пусть оценивается модель
$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u$, причем мы знаем, что $X_4 = X_2 + X_3$, тогда получим:

$$Y = \beta_1 + (\beta_2 + \beta_4) X_2 + (\beta_3 + \beta_4) X_3 + u$$

То есть невозможно отделить влияние X_4 от X_2 и X_3.

В такой ситуации, как мы рассмотрели, проблему мультиколлинеарности можно решить, просто исключив из модели X_4.

Однако чаще встречается ситуация, когда между регрессорами существует не функциональная, а стохастическая, и притом довольно тесная связь.

Тогда речь идет не о точной мультиколлинеарности, а о квазимультиколлинеарности, но ее обычно тоже называют просто мультиколлинеарностью.

В этом случае мы сможем вычислить оценки параметров модели $b = \hat{\beta} = (X'X)^{-1} X'Y$, и эти оценки по-прежнему будут несмещенными и эффективными, но оценки дисперсии будут большими, а, следовательно, оценки параметров регрессии будут незначимыми.

Проявления и последствия мультиколлинеарности

Как распознать?

- В целом построенная регрессия адекватна, коэффициент детерминации часто бывает очень высоким, но при этом коэффициенты при многих факторах незначимы.
- Вопреки априорным предположениям знаки коэффициентов противоположные, сами коэффициенты неадекватны по величине.
- Оценки параметров регрессии неустойчивы, то есть удаление или добавление небольшого количества наблюдений приводит к существенному изменению параметров.

Какие последствия?

- Возможная незначимость оценок параметров регрессии.
- Большие значения стандартных ошибок оценок параметров регрессии.
- Невозможность оценить влияние регрессоров по отдельности.

«Измерители» мультиколлинеарности

Перед началом оценивания модели можно построить матрицу парных коэффициентов корреляции между регрессорами.
Если в этой матрице присутствуют коэффициенты корреляции по модулю близкие к единице, то это может указывать на наличие мультиколлинераности.

В качестве «измерителей» мультиколлинеарности часто используют VIFы (Variance Inflation Factor):

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

В этой формуле R_j^2 – коэффициент детерминации в регрессии фактора X_j на все остальные факторы, задействованные в построении модели. Если хотя бы один из VIF_j достаточно велик (больше 8 или больше 10), то следует заподозрить мультиколлинеарность.

Также в качестве «детектора» мультиколлинеарности можно использовать собственные числа матрицы $(X'X)$. На их основе строится индекс обусловленности (conditional index):

$$CI = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$$

В этой формуле $\lambda_{\max}$ $и$ $\lambda_{\min}$ соответственно максимальное и минимальное собственные значения матрицы $(X'X)$.

Если $CI > 30$, то это может свидетельствовать о наличии мультиколлинеарности.

Что же с ней в итоге делать?!

- Некоторые исследователи считают – ничего, поскольку оценки параметров являются несмещенными и эффективными. Иногда даже прогноз по такой модели получается приемлемым.
- Добавить дополнительные наблюдения. К сожалению, это далеко не всегда возможно.
- Просто исключить из модели ту переменную, которая наиболее тесно связана с наибольшим количеством других факторов. Это приведет к некоторой потери информации, но зато оценки параметров при оставшихся переменных будут статистически значимы.
- Переопределить переменные, например, перейти к первым разностям, либо к логарифмам, и т.п.

- Использовать априорные предположения о взаимосвязи между регрессорами, т.е. фактически наложить ограничения на параметры. Тогда способ действия: см. тему о линейных ограничениях на параметры модели.
- Перейти в пространство факторов меньшей размерности. Одним из способов снижения размерности является метод главных компонент (МГК). В этом методе из исходных факторов определенным образом строятся линейные комбинации – главные компоненты, которые ортогональны. Затем построенные ГК можно использовать для построения регрессионной модели. Недостаток: ГК не всегда интерпретируемы.

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

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

Регрессия с использованием обоих факторов

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,97365166
R-квадрат	0,947997554
Нормированный R-квадрат	0,933139712
Стандартная ошибка	0,077828371
Наблюдения	10

Дисперсионный анализ

	df	*SS*	*MS*	*F*
Регрессия	2	0,772960611	0,386480306	63,8045266
Остаток	7	0,042400787	0,006057255	
Итого	9	0,815361398		

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	0,608455964	4,899733374	0,124181444	0,904662348
ln(wealth)	0,108048337	2,12638751	0,050813098	0,960894015
ln(income)	0,643406114	2,137046402	0,301072599	0,772105453

В качестве примера неустойчивости модели уберем последнее наблюдение:

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,969623225
R-квадрат	0,940169199
Нормированный R-квадрат	0,920225598
Стандартная ошибка	0,082758153
Наблюдения	9

Дисперсионный анализ

	df	*SS*	*MS*	*F*
Регрессия	2	0,645734554	0,322867277	47,14139759
Остаток	6	0,041093471	0,006848912	
Итого	8	0,686828025		

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	-0,345546124	5,64916674	-0,061167627	0,953212324
ln(wealth)	0,489127282	2,423482597	0,201828263	0,84671998
ln(income)	0,276878945	2,422324616	0,114302989	0,912726625

Коэффициенты изменились практически до неузнаваемости!

Вспомогательная регрессия

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,999514773
R-квадрат	0,999029781
Нормированный R-квадрат	0,998908504
Стандартная ошибка	0,012940484
Наблюдения	10

Дисперсионный анализ

	df	*SS*	*MS*	*F*
Регрессия	1	1,379430484	1,379430484	8237,563103
Остаток	8	0,001339649	0,000167456	
Итого	9	1,380770133		

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	2,298746873	0,056282337	40,84313082	1,42163E-10
ln(income)	1,004525017	0,011067802	90,76102194	2,42385E-13

расчет VIF **1030,695388**

Строим регрессию только на одном из факторов

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,973641809
R-квадрат	0,947978373
Нормированный R-квадрат	0,94147567
Стандартная ошибка	0,0728152
Наблюдения	10

Дисперсионный анализ

	df	*SS*	*MS*	*F*
Регрессия	1	0,772944971	0,772944971	145,7821949
Остаток	8	0,042416427	0,005302053	
Итого	9	0,815361398		

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	0,856831741	0,316696774	2,705527216	0,026843321
ln(income)	0,751943372	0,062277747	12,07402977	2,04568E-06