



ФКН

Департамент больших данных и
информационного поиска

Москва 2025

Лекция 6

Линейная регрессия: проверка условий Г-М

Машинное обучение в цифровом продукте

Полякова И.Ю.

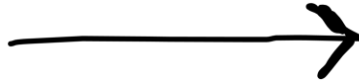
Напоминание: условия Г-М:

1. Модель правильно специфицирована;
2. Объясняющие переменные линейно независимы;
3. Ошибки независимы друг от друга;
4. Дисперсия ошибок одинакова;
5. Ошибки не зависят от наблюдений;
6. Мат. ожидание ошибок равно нулю.

Зачем требовать выполнения?



Условия Гаусса
- Маркова



Оценки
параметров
являются **BLUE**

Проверка Гауссовости ошибок

- На самом деле, не является одним из условий Г-М, но зачастую дополнительно вводится для корректного использования известных нам стат. тестов на малых выборках

Критерий Колмогорова-Смирнова

Критерий Крамера-Мизеса

Критерий Андерсона-Дарлинга

Критерий Жарка-Бера (в statsmodels называется Omnibus)

Критерий Шапиро-Уилка

QQ-plot

...

Проверка Гауссовости ошибок

Что, если ошибки не гауссовы?

- Проверить остальные условия Г-М, отработать с ними, вернуться к проверке гауссовости;
- Если есть предположение о том, что ошибки принадлежат другому распределению – попробовать оценить ММП с другим распределением ошибок;
- Если в целом Вам кажется, что все сделано правильно и данных много (настолько, что включается ЦПТ) – работать с тем, что есть

Линейная независимость

Мультиколлинеарность – наличие линейной зависимости
между регрессорами



Полная

Строгая (идеальная
линейная зависимость)



Неполная

Нестрогая (примерная
линейная зависимость)

Линейная независимость

Как проверить?

- Проверьте, что Вы не попались в ловушку дамми-переменных!
Например: в модели нет одновременно двух бинарных переменных, отвечающих за женский и мужской пол
- Постройте корреляционную матрицу
Эмпирическое правило: если есть модули корреляции больше 0.8, то проблема нестрогой мультиколлинеарности есть
- Аналог: посмотрите на определитель матрицы $X^T X$
Если он близок к нулю, то проблема нестрогой мультиколлинеарности есть

Линейная независимость

Как еще проверить?

VIF (Variance Inflation Factor) – коэффициент вздутия дисперсии

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

- Выбираем объясняющую переменную;
- Строим регрессию для этой переменной, где она сама выступит зависимой, а все остальные переменные – независимыми;
- Считаем R^2 получившейся регрессии (он и используется в формуле выше)

Эмпирическое правило:

$VIF = 1$ – отсутствие мультиколлинеарности

$1 < VIF < 5$ – умеренная мультиколлинеарность

$5 < VIF < 10$ – высокая мультиколлинеарность

$VIF > 10$ – критическая мультиколлинеарность

Линейная независимость

Как побороться с линейной зависимостью?

1. Feature engineering

- Удалить одну из коррелирующих переменных (обычно не очень хорошо);
- Изменить спецификацию на основе здравого смысла, feature importance и тд...
- Создать новые признаки из старых, понизив размерность (PCA)

2. Регуляризация

- LASSO, Ridge, ElasticNet
- Однако, быть осторожными, регуляризация делает оценки смещенными!

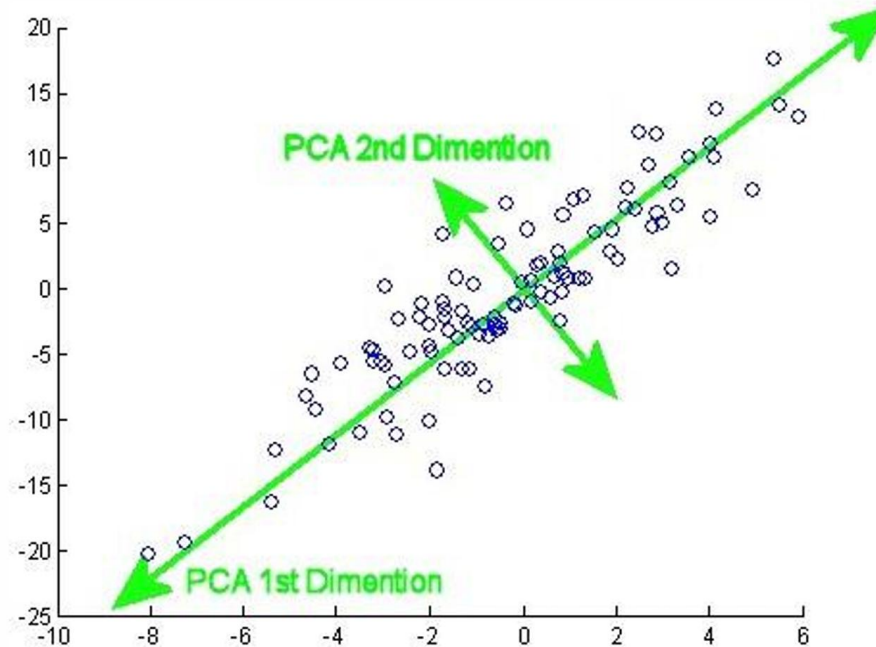
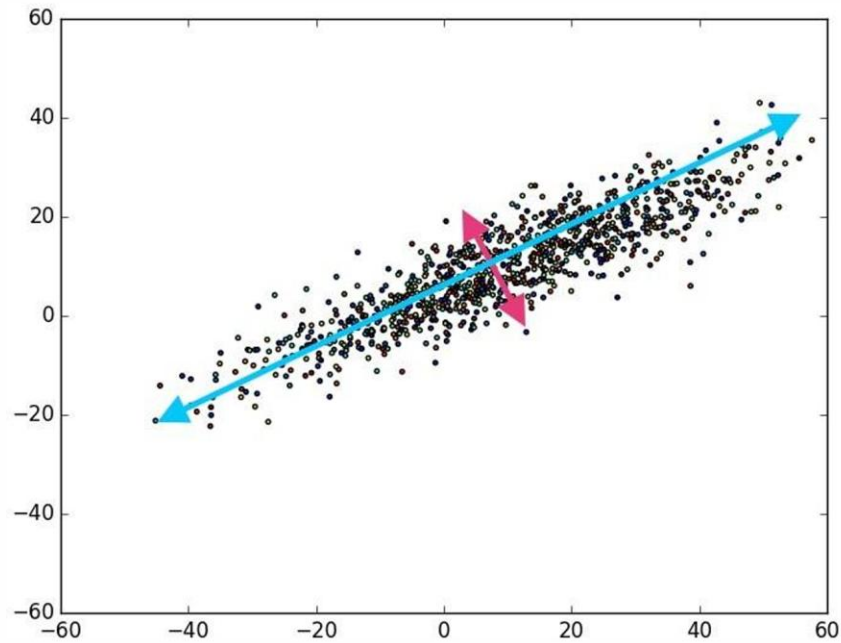
Метод главных компонент

Principal Components Analysis (PCA)

Алгоритм:

1. Исходные данные нормализуются;
2. Вычисляются собственные векторы и собственные значения матрицы ковариации данных. Эти векторы указывают направления, вдоль которых данные варьируются наиболее сильно;
3. Данные проецируются на несколько первых главных компонент. Выбор количества компонент определяется долей объяснённой дисперсии, которую они покрывают.

Метод главных компонент



Ищем такое направление, после проекции на которое, мы сохраним максимальное количество информации (дисперсии)

Метод главных компонент

- Матричная запись:

$$Z = XW^T$$

Z — матрица главных компонент

- j -й столбец W — коэффициенты при исходных признаках для вычисления нового j -го признака

$$\begin{cases} \sum W_j^T X^T X W_j \rightarrow \max_W \\ W^T W = 1 \end{cases}$$

Новые признаки после PCA являются линейной комбинацией из старых признаков

Метод главных компонент

Пример: из 21й сенсорной характеристики в 8

Table.1 The results of common factor analysis

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8
Color intensity								
Odor intensity								
Taste intensity						0.567		
Foreign tastes		-0.709						
Crystallization								0.427
Florality				0.932				
Fruitiness			0.963					
Berry			0.604					
Herbivory								
Woodiness	0.438							
Spice	0.620							
Wax								0.679
Artificial sugar		-0.567						
Sweetness						0.678		
Bitterness	0.464			0.458				
Sourness							0.438	
Peppercorn				0.894				
Aftertaste	0.653					0.466		
Honey plant		0.495						
Tartness	0.764							
Astringency	0.530							

1. Tartness
2. Absence of extraneous odors and components
3. Fruit and berry flavor
4. Peppercorn (the tickling sensation in the throat after the honey tasting)
5. Florality
6. Intensity of sweet flavor
7. Sourness
8. Sensation of taste, odor, or texture of the wax.

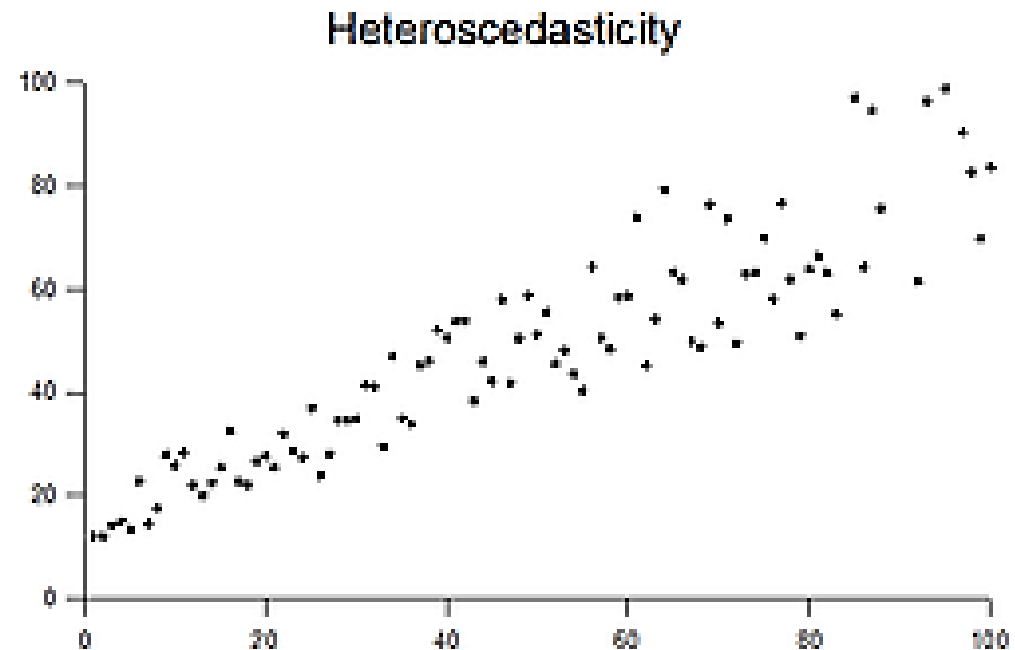
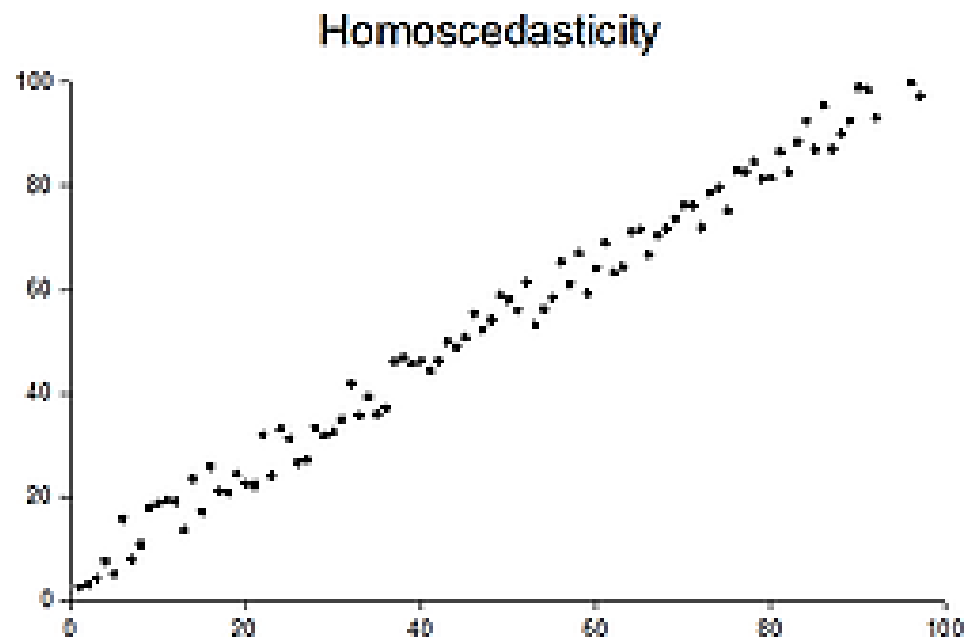


Источник:

Zaripova J., Chuprianova K. C., Polyakova I., Semenova D., Kulikova S. [The impact of sensory characteristics on the willingness to pay for honey.](#) 2023.

Гомоскедастичность

Проблема: дисперсия ошибок непостоянна/по главной диагонали ковариационной матрицы ошибок находятся разные числа



Источник: https://en.wikipedia.org/wiki/Homoscedasticity_and_heteroscedasticity

Гомоскедастичность

Проблема: дисперсия ошибок непостоянна/по главной диагонали ковариационной матрицы ошибок находятся разные числа

$$H_0: \sigma^2 I = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}$$

Проверять равенство всех дисперсий в ковариационной матрице не вариант, никогда не хватит данных

Как тогда проверить?

Гомоскедастичность

Критерий Гольдфельда-Куандта

Алгоритм:

Предпосылка: $\sigma^2 \sim \text{признак}$

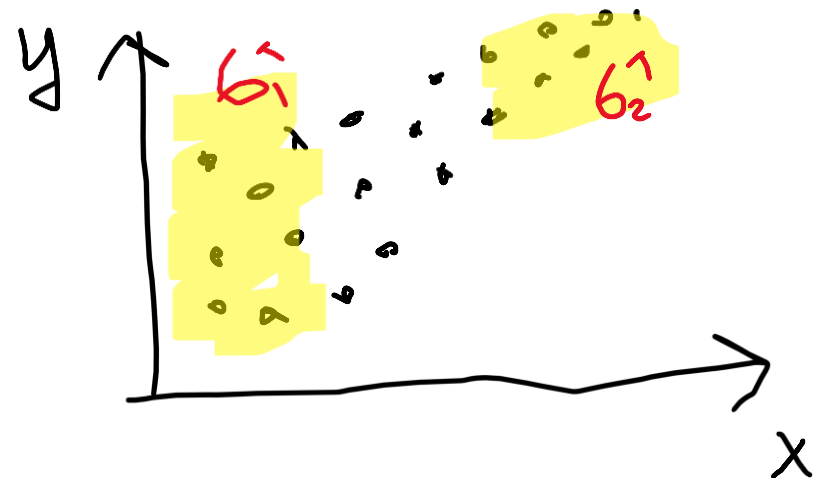
1. Сортируем данные по x (выбранный признак);
2. Из середины ряда убираем d наблюдений;
3. Строим линейную регрессию в двух оставшихся крайних группах;
4. Сравниваем суммы квадратов остатков

$$\hat{\sigma}^2 = \sum \hat{\varepsilon}_i^2 = \sum (y_i - (X\theta)_i)^2$$

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

$$T = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \sim F(n_1 - k_1; n_2 - k_2)$$



Гомоскедастичность

Критерий Бреуша-Пагана

Предпосылка: дисперсия остатков зависит от совокупности факторов

Алгоритм:

1. Строим неограниченную регрессию, считаем ее квадраты остатков;
2. Строим ограниченную регрессию на квадраты остатков с p признаков;
3. Считаем RSS ограниченной модели.

$$T = \frac{RSS_R}{2} \sim \chi^2_p$$

Гомоскедастичность

Критерий Уайта

Предпосылка: дисперсия остатков зависит от совокупности факторов

Алгоритм:

1. Строим неограниченную регрессию, считаем ее квадраты остатков;
2. Строим регрессию на квадраты остатков, где регрессорами выступают исходные признаки, их квадраты и попарные произведения;
3. Считаем RSS вспомогательной модели.

$$H_0 = \lambda_1 = \dots = \lambda_m = 0 \quad \text{гомоск.}$$

$$H_1 = \exists \lambda_i \neq 0 \quad \text{гетероск.}$$

$$T = n \cdot R^2_{\text{всп.}} \sim \chi^2_m$$

Гомоскедастичность

Что делать, если все-таки обнаружили гетероскедастичность?

1. Преобразовать зависимую переменную (часто: логарифмирование, извлечение квадратного корня);
2. Использовать робастные ошибки (ошибки в форме Уайта): оставляем несмещенные оценки МНК, но изменяем формулы расчета стандартных ошибок с учетом гетероскедастичности. Данный перерасчет ошибок рекомендуют **использовать почти всегда**;
3. Использовать взвешенный МНК: предполагаем, как именно меняется дисперсия. Присваиваем каждому наблюдению вес, обратно пропорциональный его дисперсии;
4. Изменить спецификацию модели: добавить новые признаки, полиномиальные признаки.

Гомоскедастичность

Робастные ошибки (ошибки в форме Уайта)

Являются состоятельными при гетероскедастичности

Гомоскедастичность:

$$\text{var}(\hat{\theta}) = \hat{\sigma}^2 (X^T X)^{-1}, \text{ где } \hat{\sigma}^2 = \frac{RSS}{n-k}$$

Гетероскедастичность
(при отсутствии
автокорреляции):

$$\text{var}(\hat{\theta}) = (X^T X)^{-1} \left(\sum_{t=1}^n e_t^2 x_t x_t^T \right) (X^T X)^{-1}$$

Вместо неизвестных дисперсий – используем
оцененные квадраты ошибок

Стандартные ошибки в форме Уайта никак не влияют на расчет самих оценок модели, однако позволяют скорректировать доверительные интервалы и значимость коэффициентов

Автокорреляция

Проблема: ошибки разных наблюдений связаны между собой /
ошибка одного наблюдения «тянет» за собой ошибку другого

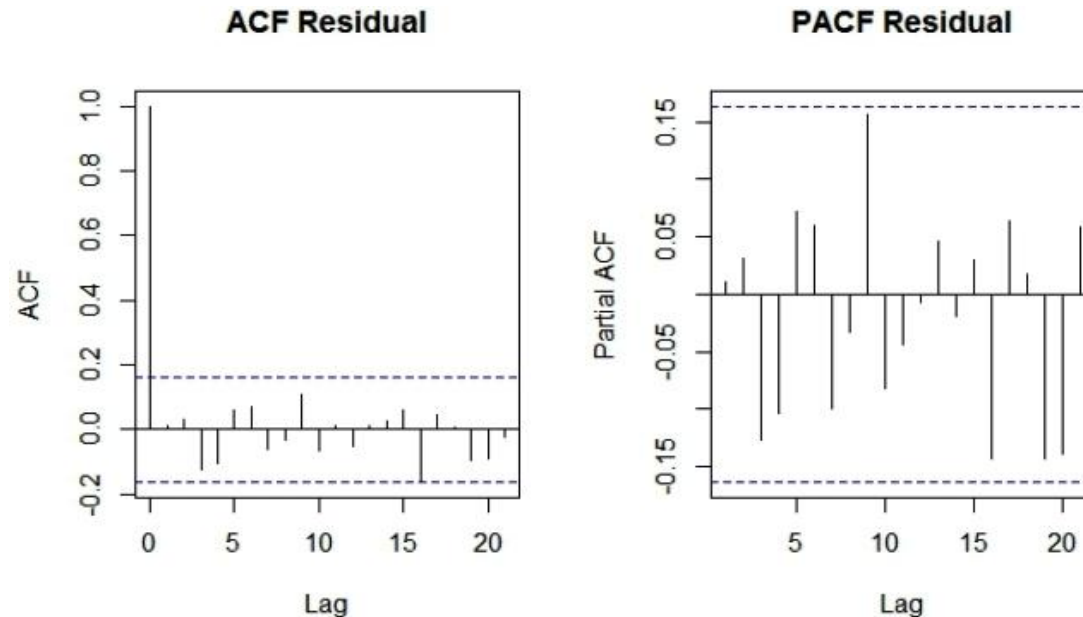
$$\text{COV}(\varepsilon_i, \varepsilon_j) \neq 0 \quad i \neq j$$

Автокорреляция, гетероскедастичность и эндогенность часто ходят «вместе», возможно, если исправите что-то одно (скорее всего, речь пойдет про эндогенность), то есть приличный шанс, что остальное «сгладится»

Автокорреляция

Как увидеть?

1. Нарисовать график остатков – если похоже на случайный шум, то автокорреляции нет. Если есть «паттерн» - скорее всего, есть. Более того, обычно автокорреляция автоматически возникает при наличии гетероскедастичности;
2. Построить ACF (график автокорреляции) и PCF (график частичной автокорреляции), посмотреть, есть ли большие корреляции по модулю (которые превосходят дов. интервал для нуля)



Автокорреляция

Критерий Дарбина-Уотсона

Предпосылка: $\varepsilon_t = \rho \varepsilon_{t-1} + \mu_t$ iid $N(0, \sigma^2)$

$H_0: \rho = 0$ Автокорреляции нет

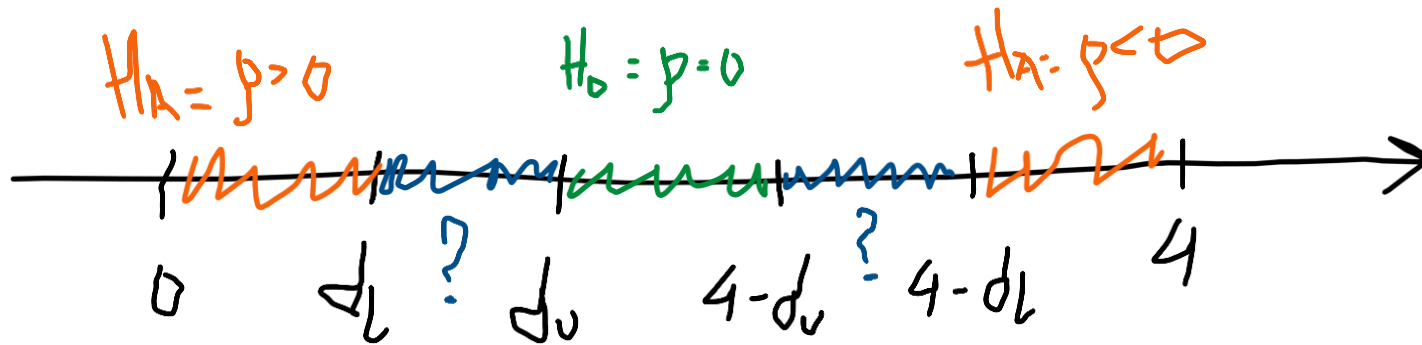
$H_A: \rho \neq 0$ Автокорреляция есть

$$T = \frac{\sum_{t=2}^n (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^n \hat{\varepsilon}_t^2} \in [0; 4]$$

Автокорреляция

Критерий Дарбина-Уотсона

d_L, d_U – Квантили распределения Дарбина-Уотсона



Тест Дарбина-Уотсона не совсем классический в плане принятия решения, так как есть «зоны неопределенности»

Подробнее:

1. https://en.wikipedia.org/wiki/Durbin%E2%80%93Watson_statistic
2. Durbin, J.; Watson, G. S. (1971). "Testing for serial correlation in least squares regression.III". *Biometrika*. **58** (1): 1-19. doi:[10.2307/2334313](https://doi.org/10.2307/2334313). JSTOR [2334313](https://www.jstor.org/stable/2334313)

Автокорреляция

Критерий Льюинга-Бокса

Рассчитаем корреляции вида:

$$\rho_p = \text{corr}(\varepsilon_{t+p}, \varepsilon_t)$$

$$H_0 = \rho_1 = \rho_2 = \dots = \rho_p = 0$$

$$H_A = \exists \rho_j \neq 0$$

Автокорреляции нет

Автокорреляция есть

$$T = n(n+2) \sum_{j=1}^p \frac{\hat{\rho}_j^2}{n-j} \sim \chi_p^2$$

Гомоскедастичность и автокорреляция

Робастные ошибки (ошибки в форме Ньюи-Уэста)

Являются состоятельными при гетероскедастичности и автокорреляции

Гомоскедастичность: $\text{var}(\hat{\theta}) = \hat{\sigma}^2 (X^T X)^{-1}$, где $\hat{\sigma}^2 = \frac{RSS}{n-k}$

Гетероскедастичность (с автокорреляцией):

$$\text{var}(\hat{\theta}) = (X^T X)^{-1} \left(\sum_{t=1}^n e_t^2 x_t x_t^T + \sum_{j=1}^L \sum_{t=j+1}^n w_j e_t e_{t-j} (x_t x_{t-j}^T + x_{t-j} x_t^T) \right) (X^T X)^{-1}$$

Кроме главной диагонали в ковариационной матрице ошибок требуется оценить еще побочные элементы (есть нюансы с подбором параметров)

Эндогенность

Одна или несколько объясняющих переменных (X) скоррелированы с ошибкой (ϵ) в модели регрессии

Причины:

1. Пропущенная переменная: переменная, скоррелированная с пропущенной «вбирает» в себя эффект пропущенной;
2. Одновременность: эффект смещен, так как влияние двух переменных взаимное;
3. Ошибки измерения: переменная, измеренная с ошибкой, содержит в себе ошибку и оказывается скоррелированной с остатками (более того, чаще всего коэффициент при переменной с ошибкой стремится к нулю)

Тест на пропущенную переменную

RESET-тест (тест Рамсея):

Предпосылка:

В модели есть пропущенные переменные, если это так, то «корректировка модели самой себя» даст прирост объясненной дисперсии

Алгоритм:

1. Оцениваем базовую модель, получаем предсказания;
2. Оцениваем модель где помимо стандартных регрессоров будут содержаться степени (до $p-1$) предсказаний базовой модели;
3. Проверяем гипотезу о том, что все множители перед полиномами предсказаний равны нулю.

$$H_0 = \lambda_1 = \dots = \lambda_p = 0$$

$$H_1 = \exists \lambda_i \neq 0$$

$$T \sim F_{p, n-k-p}$$

Статистика
аналогична F-тесту с
моделью с
ограничениями и
без ограничений

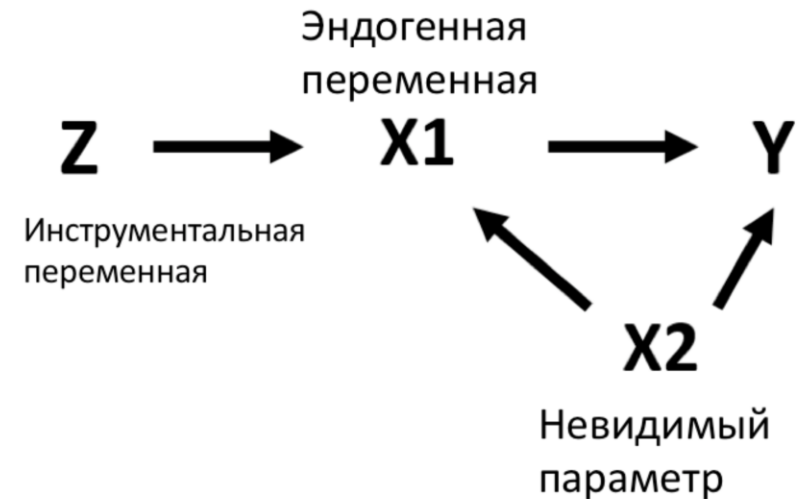
Инструментальные переменные

Как бороться с эндогенностью?

Идея: найдем прокси-переменную/ые для той, которую пропустили

Какими должны быть инструменты:

1. Инструмент экзогенен (не коррелирует с остатком);
2. Инструмент значим (сильно коррелирует с эндогенной переменной)




Инструментальные переменные

Двухшаговый МНК:

Первый шаг: регрессируем эндогенную переменную на экзогенные переменные и инструменты;

Второй шаг: регрессируем зависимую переменную на все экзогенные переменные и предсказанную эндогенную переменную с первого шага.



При корректно
выбранных
инструментах
оценки
состоятельны

Резюме по условиям Г-М

Нарушенное условие	Последствия
Неполная линейная зависимость	Неэффективные оценки, доверительные интервалы шире истинных
Автокорреляция	Неэффективные оценки, доверительные интервалы уже истинных
Гетероскедастичность	Неэффективные оценки, стандартные ошибки оценок смещены, доверительные интервалы ненадежны
Эндогенность	Оценки смещены, неэффективны, несостоятельны

Условие про нулевое мат. ожидание ошибок зачастую автоматически выполнено при включении константы в модель

Дополнительно

- Интересно про применение критериев согласия и выбор распределения по данным:
<https://habr.com/ru/companies/tbank/articles/911900/>
- Про метод главных компонент: <https://habr.com/ru/articles/304214/>
- Лекции Ольги Демидовой по эконометрике, в частности здесь про эндогенность и инструменты:
<https://www.hse.ru/mirror/pubs/share/565769150.pdf>
- Лекции Бориса Демешева по эконометрике:
https://www.youtube.com/watch?v=Ucwl7tY7bss&list=PLu5flfwrnSD5d02G9YJcDv30Fp5_70-sl (есть отдельный плейлист про эндогенность)

