

ЛЕКЦИЯ 13. Модельные допущения и нормальность остатков: на чём держится линейная регрессия


Введение

Вы уже научились многому:

- строить разные типы регрессий,
- использовать регуляризацию,
- сравнивать модели,
- интерпретировать коэффициенты и остатки.

Но всё это работает **только при одном условии:**

если **допущения модели** соблюдены.

 Допущения — это **невидимые правила**, которые лежат в основе линейной регрессии.

Если они нарушаются — модель может **давать ложные коэффициенты**, **переворачивать выводы**, или выглядеть “красиво” снаружи, но “лженаучно” внутри.

Основные допущения линейной регрессии

№	Допущение	Что это значит
1	Линейность	Связь между X и Y — линейная
2	Независимость наблюдений	Данные не зависят друг от друга (например, не повторяются)
3	Постоянная дисперсия (гомоскедастичность)	Остатки равномерны по всей шкале
4	Нормальность остатков	Остатки распределены нормально (\approx колокол)

5	Отсутствие мультиколлинеарности	Переменные не дублируют друг друга
---	---------------------------------	------------------------------------

Почему важно?

Если вы нарушаете допущения:

- t - и F -тесты становятся недостоверными
- коэффициенты могут быть переоценены
- доверительные интервалы — неверны
- визуально модель “вроде работает”, но **вы делаете неправильные выводы**



Сегодня сосредоточимся на нормальности остатков

Что значит “нормальность остатков”?

Остатки ($e = y - \hat{y}$) должны:

- быть **распределены случайно**,
- **не иметь смещения**,
- образовывать график в форме **нормального распределения** (колокол).

Это особенно важно для:

- расчёта доверительных интервалов
- корректности t - и F -статистики
- построения прогнозов

Как проверить нормальность остатков

1. Графически (гистограмма или QQ-график)

- Постройте гистограмму остатков
- Если она похожа на колокол → ок
- Если есть перекосяк, “горб”, скошенность → проблема

2. Статистически (если есть Python / R)

- Тест Шапиро–Уилка
- Jarque–Bera
- Anderson–Darling

Если $p\text{-value} < 0.05 \rightarrow$ остатки не нормальны

Пример визуального анализа остатков

Остатки выглядят как:	Интерпретация
Симметричный колокол	ОК, нормальность соблюдена
Скошены вправо или влево	Нарушена нормальность
Имеют два “пика”	Вероятно, пропущен важный фактор
В виде “ступеней” или “провалов”	Ошибка спецификации

Что делать, если нормальность нарушена?

Проблема	Решение
Остатки скошены	Попробуйте логарифмировать Y ($\ln(y)$)
Остатки с “двойным пиком”	Проверьте: не пропущен ли категориальный фактор (например, пол?)
Модель “жмёт” часть данных	Постройте полиномиальную модель или разделите данные на группы

Логика проверки

1. Постройте модель
2. Постройте график остатков
3. Проверьте форму распределения
4. При необходимости — трансформируйте переменные
5. Снова проверьте остатки

Использование ИИ

Инструмент	Что делает
------------	------------

ChatGPT	Объяснит, нарушено ли распределение остатков
Excel Copilot	Построит гистограмму и предложит вывод
Notion AI	Поможет сформулировать анализ ошибки модели

Запрещено:

- Игнорировать форму остатков
- Интерпретировать модель, если остатки систематически нарушают нормальность
- “Подгонять” график вручную
- Не фиксировать допущения в исследовательском отчёте

Вывод

Модель — это не просто “формула с R^2 ”.

Это система, работающая **только при соблюдении правил**.

Нормальность остатков — один из фундаментальных принципов статистической корректности.

Если вы научились замечать, когда модель **врет в остатках**, вы стали настоящим аналитиком.