

Лекция 6. Гетероскедастичность

Введение: от идеального мира к реальному

До сих пор мы работали в рамках классической линейной модели, где дисперсия ошибок постоянна ($Var(\varepsilon_i) = \sigma^2$). Это предположение называется гомоскедастичность. Однако в реальных данных это часто не так.

Гетероскедастичность — это нарушение предположения о постоянстве дисперсии случайных ошибок. Когда $Var(\varepsilon_i) = \sigma^2$, то есть дисперсия ошибки изменяется от наблюдения к наблюдению.

Аналогия: представьте, что вы предсказываете расходы семьи. Для семей с низким доходом расходы довольно предсказуемы (маленький разброс вокруг среднего). Для богатых семей разброс огромен: одна может потратить деньги на яхту, другая - вложить в бизнес. Дисперсия ошибки предсказания зависит от уровня дохода.

1. Что такое гетероскедастичность?

Гомоскедастичность: Разброс точек вокруг линии регрессии примерно одинаков для всех значений X (Рисунок 6.1).

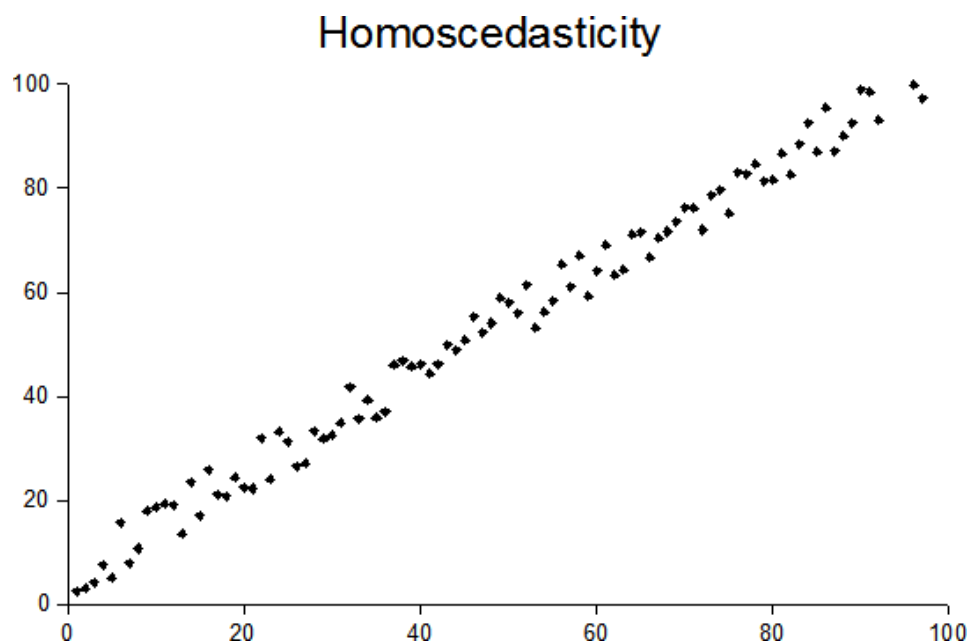


Рисунок 6.1. Гомоскедастичность

Гетероскедастичность: Разброс точек изменяется с изменением X . Часто имеет форму "веера" или "конуса" (Рисунок 6.2).

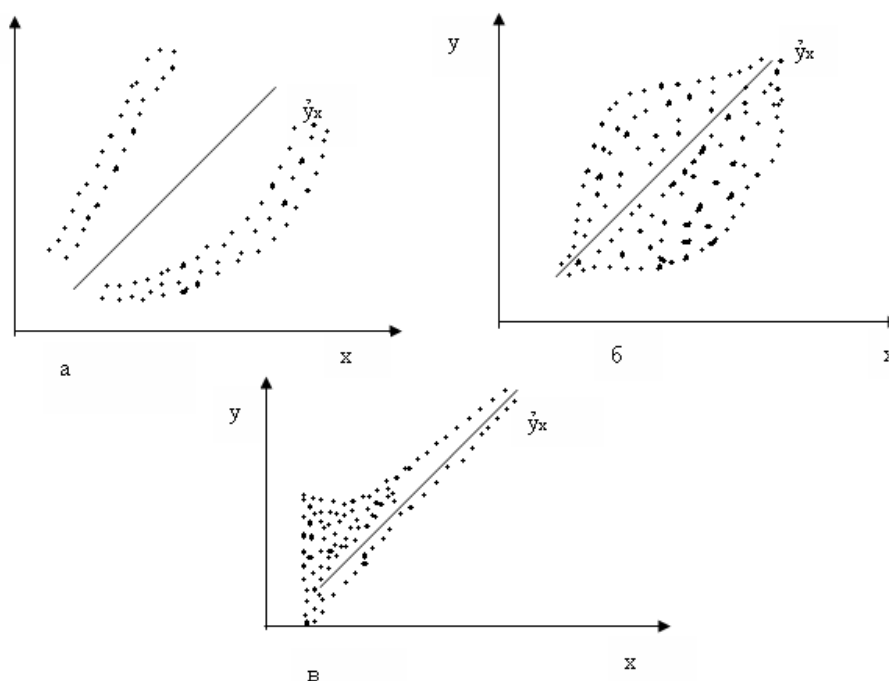


Рисунок 6.2. Гетероскедастичность

Формальное определение: Нарушение 4-й предпосылки КЛМР:

$$Var(\varepsilon_i|X) = \sigma_i^2,$$

где $\sigma_i^2 \neq \sigma_j^2$ для некоторых $i \neq j$.

2. Причины и последствия гетероскедастичности

2.1. Типичные причины:

- **Природа данных:** в данных есть "единицы" разного масштаба (например, наблюдения за малыми и крупными фирмами).

- **Эффект обучения:** стечением времени ошибки прогноза уменьшаются (временные ряды).
- **Пропуск важной переменной.**
- **Ошибки измерения в переменных.**

2.2. Последствия для МНК-оценок:

1. **Оценки коэффициентов (b) остаются несмещенными и состоятельными.** Это хорошая новость!
2. **Стандартные ошибки коэффициентов становятся смещенными.** Это очень плохая новость!
 - Обычные формулы $se(b)$ неверны.
 - **t-статистики** и **F-статистики** теряют свою распределения (t и F).
 - **Доверительные интервалы** строятся неверно.
 - **Проверка гипотез** становится ненадежной.

Итог: Мы можем делать неверные выводы о значимости коэффициентов, даже если сами коэффициенты оценены правильно.

3. Обнаружение гетероскедастичности

3.1. Графические методы

- **Анализ остатков:** строим график остатков e_i против предсказанных значений \hat{Y}_i или против одной из объясняющих переменных.
- **Что ищем?** Любую систематическую форму (веер, конус, U-образную форму), а не просто случайное облако (Рисунок 6.3).

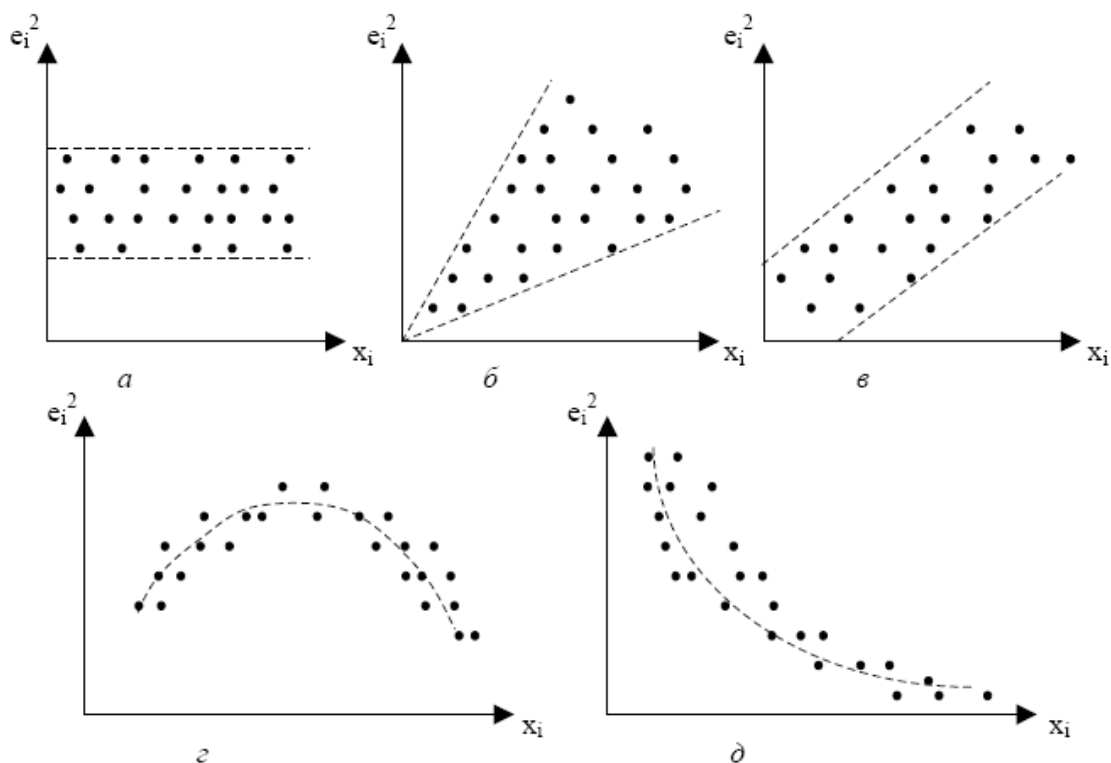


Рисунок 6.3. График остатков при гетероскедастичности

3.2. Статистические тесты

А. Тест Уайта (White Test) - универсальный

- **Идея:** проверить, зависит ли дисперсия ошибок от объясняющих переменных, их квадратов и попарных произведений.
- **Шаги:**
 1. Оцениваем исходную регрессию, получаем остатки e_i .
 2. Оцениваем вспомогательную регрессию: $e_i^2 = \delta_0 + \delta_1 X_1 + \delta_2 X_2 + \dots + \delta_3 X_1^2 + \delta_4 X_2^2 + \delta_5 X_1 X_2 + \dots + u_i$.
 3. Нулевая гипотеза H_0 : **гомоскедастичность** (все δ во вспомогательной регрессии, кроме константы, равны 0).
 4. Используем статистику $LM = n * R_2^2 \sim \chi_n^2$, где n - число регрессоров во вспомогательной регрессии (без константы).
 5. Если $LM > \chi^2_{\text{крит}}$ или $p\text{-value} < \alpha$, **отвергаем H_0** — есть гетероскедастичность.

В. Тест Бреуша-Пагана (Breusch-Pagan Test) — более простой

- **Идея:** Аналогична тесту Уайта, но во вспомогательной регрессии используются только исходные регрессоры (без квадратов и произведений).

$$e_i^2 = \delta_0 + \delta_1 X_1 + \delta_2 X_2 + \dots + u_i$$

- Более мощный, если форма гетероскедастичности простая, но может не уловить сложные зависимости.

4. Методы борьбы с гетероскедастичностью

4.1. Робастные стандартные ошибки (ошибки Уайта)

Самое популярное и простое решение. Мы не меняем МНК-оценки коэффициентов (β), но **пересчитываем** их стандартные ошибки так, чтобы они были состоятельными при наличии гетероскедастичности.

- **Преимущества:**
 - Коэффициенты МНК остаются BLUE (при гомоскедастичности) или несмещенными (при гетероскедастичности).
 - Не нужно знать точную форму гетероскедастичности.
 - Легко реализуется в любом статистическом пакете.
- **Результат:** Мы получаем **скорректированные t-статистики, p-value** и **доверительные интервалы**, на которые можно полагаться.

4.2. Обобщенный метод наименьших квадратов (ОМНК / GLS)

- **Идея:** Если мы знаем или можем предположить форму гетероскедастичности ($\text{Var}(\varepsilon_i) = \sigma^2 * h_i$), мы можем **преобразовать данные**, чтобы сделать дисперсию постоянной.
- **Метод взвешенных наименьших квадратов (ВМНК / WLS):**
Частный случай ОМНК. Мы делим все переменные модели на $\sqrt{h_i}$ (например, на X_i , если дисперсия пропорциональна X_i^2).

Исходная модель: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, где $\text{Var}(\varepsilon_i) = \sigma^2 X_i^2$

Преобразованная модель: $Y_i/X_i = \beta_0(1/X_i) + \beta_1 + (\varepsilon_i/X_i)$

Дисперсия новой ошибки: $\text{Var}(\varepsilon_i/X_i) = \sigma^2 \rightarrow$ гомоскедастичность!

- **Недостаток:** нужно правильно угадать форму h_i .

5. Практический пример в три шага

Данные: Зависимость расходов на еду от дохода семьи.

1. **Оцениваем модель МНК:** $\text{Расходы_на_еду} = b_0 + b_1 * \text{Доход} + e$
Получаем: $b_1 = 0.5$ ($t = 5.0$, $p < 0.001$). Коэффициент значим.
2. **Проверяем на гетероскедастичность (Тест Уайта):**
 - Строим график остатков: виден "веер".
 - Проводим тест Уайта: $LM = 15.2$, $p\text{-value} = 0.0005$.
 - **Вывод:** Гетероскедастичность присутствует.
3. **Принимаем меры (Робастные ошибки Уайта):**
 - Пересчитываем стандартные ошибки. Новая $\text{'se}(b_1)'$ больше старой.
 - Новая t-статистика: $t = 3.2$, $p\text{-value} = 0.002$.
 - **Итоговый вывод:** Коэффициент b_1 по-прежнему статистически значим на 1% уровне, но его точность (стандартная ошибка) была переоценена обычным МНК.

Резюме

1. **Гетероскедастичность** - это непостоянство дисперсии ошибок, частое явление в реальных данных.
2. **Основная проблема** - неверные стандартные ошибки и, как следствие, ненадежные статистические выводы.
3. **Обнаружить** ее можно с помощью графиков остатков и формальных тестов (Уайта, Бреуша-Пагана).
4. **Решить проблему** проще всего с помощью **робастных стандартных ошибок Уайта**, которые делают выводы корректными.

5. Более сложный метод - **ОМНК/ВМНК**, который может повысить эффективность оценок, если форма гетероскедастичности известна.

На следующей лекции: Мы изучим другую общую проблему - **автокорреляцию** в остатках, характерную для временных рядов.

Вопросы для самопроверки:

1. Почему при гетероскедастичности обычные t -тесты могут показывать "значимость" незначимых коэффициентов?
2. В чем ключевое различие между подходами Робастных ошибок Уайта и Взвешенного МНК?
3. Может ли наличие гетероскедастичности сделать МНК-оценки смещенными?