

**Обобщенная линейная модель
множественной регрессии с
гетероскедастичными остатками**

ОЛММР с гетероскедастичными остатками

Линейная модель множественной регрессии

$$Y = X\bar{\beta} + \bar{\varepsilon},$$

для которой нарушено 4 условие Гаусса-Маркова называется обобщенной линейной моделью множественной регрессии (ОЛММР) с гетероскедастичными остатками, а именно:

- 1) x_1, \dots, x_k – детерминированные переменные;
- 2) ранг матрицы X равен " $k+1$ " – среди признаков нет линейно зависимых;
- 3) $M\varepsilon_i = 0, i = \overline{1, n}$ - нет систематических ошибок в измерении y ;
- 4) $D\varepsilon_i = M\varepsilon_i^2 = \sigma_i^2, i = \overline{1, n}$
- 5) $\text{cov}(\varepsilon_i, \varepsilon_j) = M(\varepsilon_i \cdot \varepsilon_j) = 0, i \neq j, i = \overline{1, n} j = \overline{1, n}$

4') $\Sigma_{\varepsilon} = M\bar{\varepsilon}\bar{\varepsilon}^T = \sigma^2\Sigma_0$ ($\sigma_i^2 \neq \sigma_j^2$, есть хотя две различные дисперсии), т.е. на диагонали стоят неравные дисперсии, а вне диагональные элементы равны 0.

Гетероскедастичность — определение

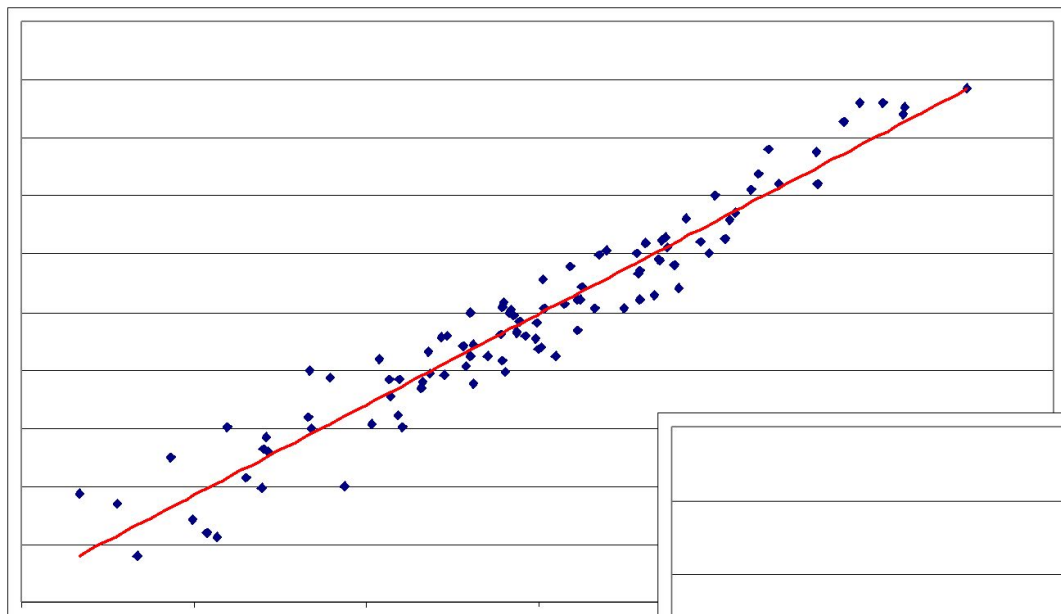
Гомоскедастичность (условие №4 теоремы Гаусса — Маркова) — случайные ошибки имеют

постоянную дисперсию: $V(\varepsilon_i) = \sigma^2 = const$

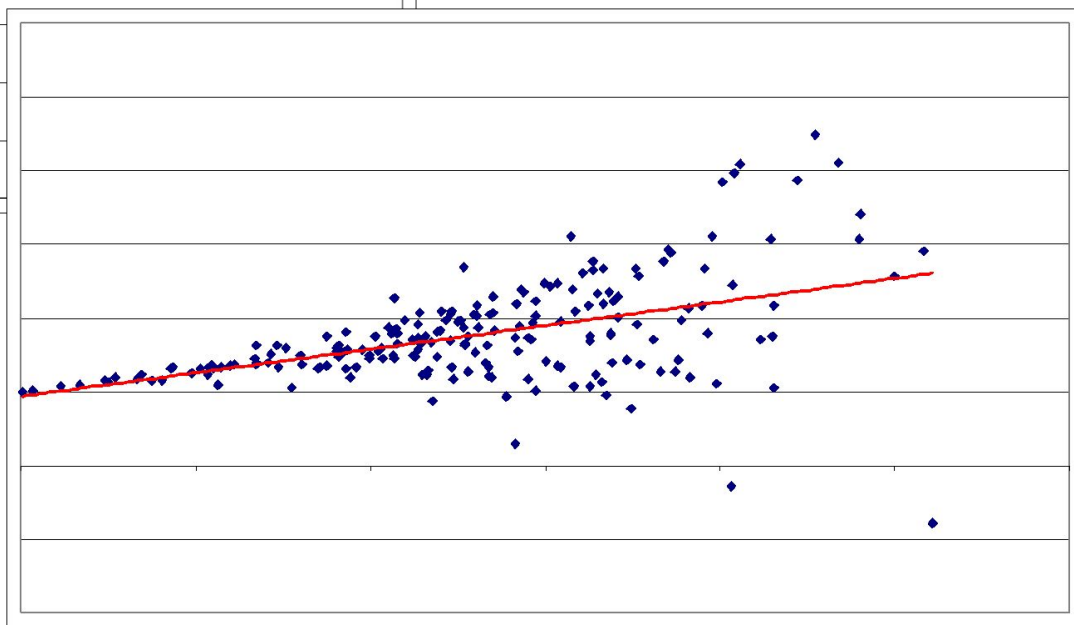
Гетероскедастичность — случайные ошибки имеют

непостоянную дисперсию: $V(\varepsilon_i) = \sigma_i^2 \neq const$

Гетероскедастичность — графическая иллюстрация

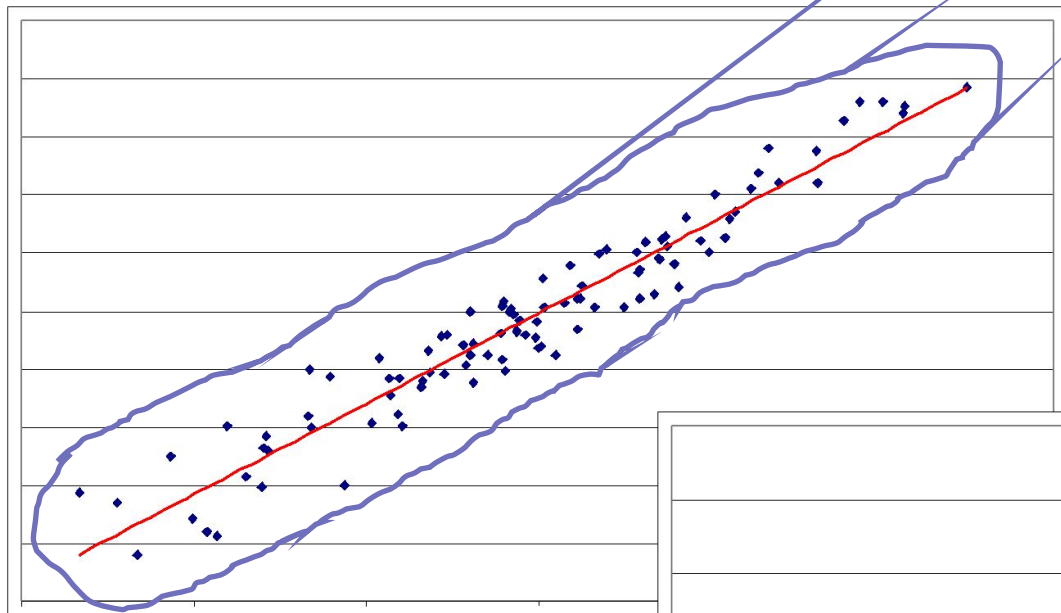


Нет
гетероскедастичности
⇐=



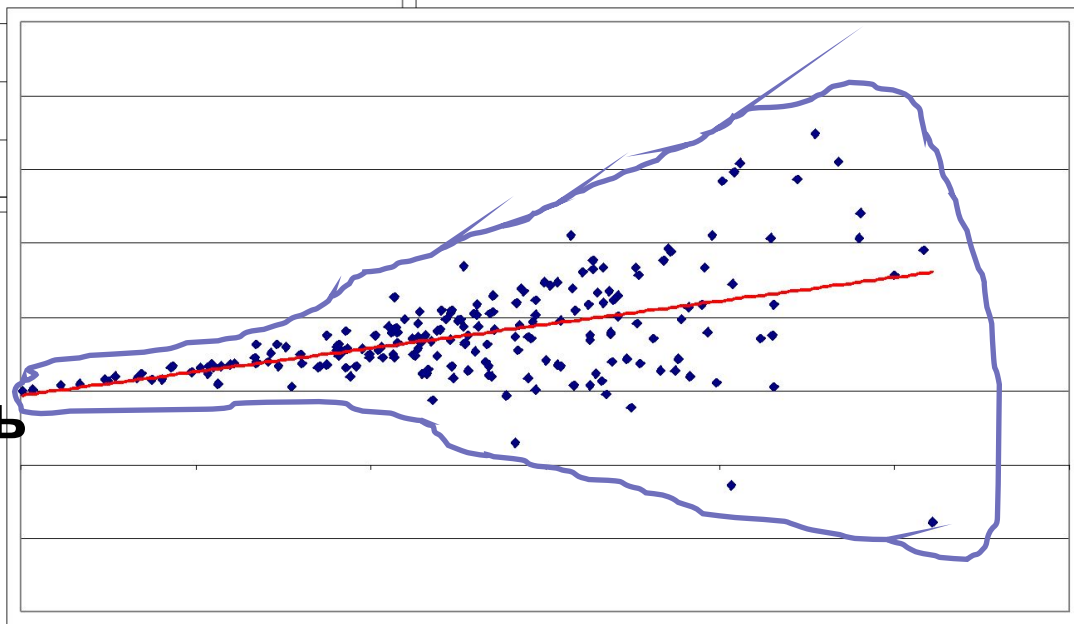
Есть
гетероскедастичность
=>

Гетероскедастичность — графическая иллюстрация

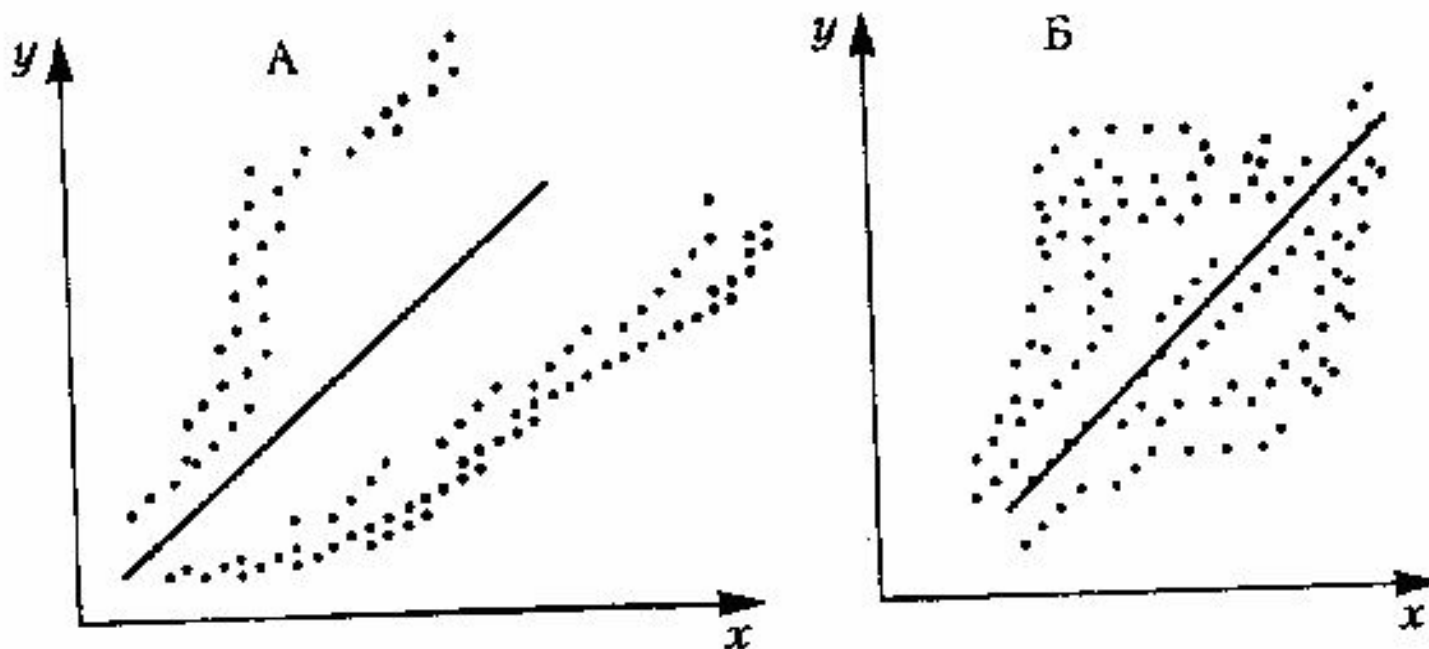


Нет
гетероскедастичности
⇐=

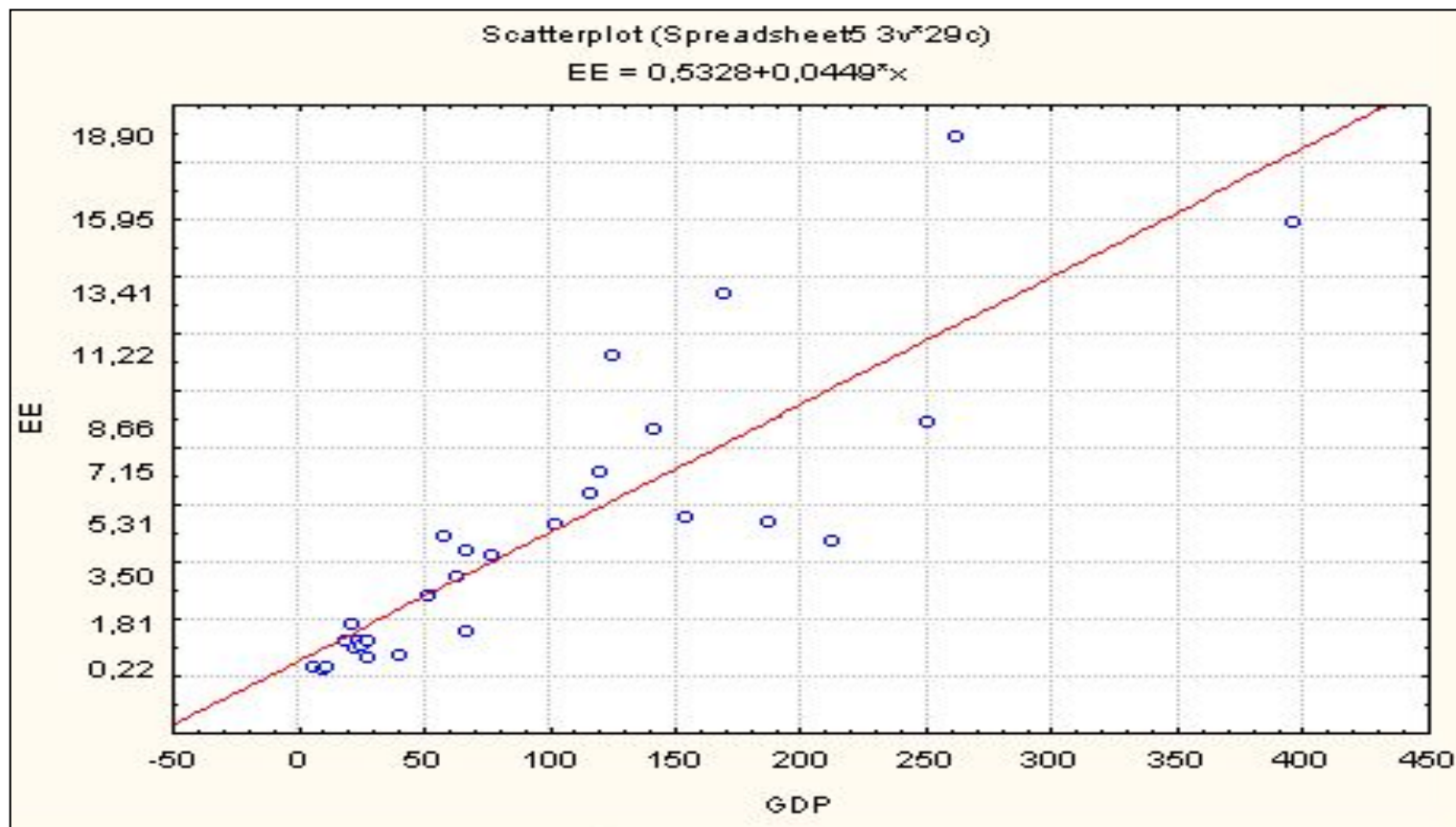
Есть
гетероскедастичность
⇒



Пример 1. При исследовании среднедушевых сбережений и дохода в семьях разброс в данных будет выше для семей с более высокими доходами. Это означает, что дисперсия зависимых величин, а, следовательно, и случайных ошибок не постоянны.



Пример 2. При изучении влияния ВВП на затраты на образование по статистическим данным странам мира разброс значений относительно функции регрессии выше у стран с более высокими значениями ВВП



Гетероскедастичность — ковариационная матрица вектора случайных ошибок

Нет гетероскедастичности:

$$\Sigma_{\varepsilon} = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

Есть гетероскедастичность:

$$\Sigma_{\varepsilon} = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix} \quad \Sigma_{\varepsilon} = \{Cov(\varepsilon_i, \varepsilon_m)\}$$

Последствия гетероскедастичности

1. МНК-оценки коэффициентов остаются несмещенными, но...
2. МНК-оценки коэффициентов становятся неэффективными.
3. Стандартные ошибки коэффициентов смещены и несостоятельны => t-статистики вычисляются некорректно.

Последствия гетероскедастичности

На практике в пространственных выборках гетероскедастичность есть (почти) всегда.

К счастью, существуют простые и действенные средства решения этой проблемы, о которых пойдет речь ниже

Последствия гетероскедастичности

На практике в пространственных выборках гетероскедастичность есть (почти) всегда.

К счастью, существуют простые и действенные средства решения этой проблемы, о которых пойдет речь ниже

Но сначала обсудим, как выявить наличие этой проблемы

Обнаружение гетероскедастичности

- Анализ графиков остатков
- Тесты на гетероскедастичность:
 - Тест ранговой корреляции Спирмена
 - Голдфелда — Квандта
 - Глейзера
 - Бреуша — Пагана
 - Уайта

Тесты для проверки гетероскедастичности

Тест ранговой корреляции Спирмена

1. Вычисляется коэффициент ранговой корреляции Спирмена

$$\rho_{x,e} = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n^3 - n},$$

где d_i — разность между рангами значений x_i и e_i

2. Выдвигается гипотеза

$$H_0 : \rho_{xe} = 0$$

$$H_1 : \rho_{xe} \neq 0$$

3. Гипотеза проверяется на основе статистики

$$|t| = \frac{|\rho_{x,e}| \sqrt{n-2}}{\sqrt{1-\rho_{x,e}^2}} \in St(n-2) / H_0$$

Тесты для проверки на гетероскедастичность

Тест Голдфелда—Квандта

1. Все n наблюдений X и Y упорядочиваются по объясняющей переменной, влиянием которой порождается гетероскедастичность;
2. Оцениваются коэффициенты уравнений регрессии для первых n' и последних n'' наблюдений, причем

$$n' = n'' \approx \frac{3}{8}n$$

3. Выдвигаются гипотезы

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_n^2 \text{ (нет гетероскедастичности)}$$

$$H_1 : \exists i \neq j : \sigma_i^2 \neq \sigma_j^2 \text{ (есть гетероскедастичность)}$$

4. Вычисляются суммы квадратов отклонений для первых n' и последних n'' наблюдений

$$Q' = (\bar{e}')^T (\bar{e}') \quad Q'' = (\bar{e}'')^T (\bar{e}'')$$

Тест Голдфелда—Квандта

5. Строятся статистики

$$F = \frac{\max(Q'; Q'')}{\min(Q'; Q'')}$$

6. В случае отклонения нулевой гипотезы структура матрицы Σ_0 имеет вид

Если $Q'' > Q'$

$$\Sigma_0 = \begin{pmatrix} x_{1j}^2 & 0 & 0 & 0 \\ 0 & x_{2j}^2 & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & x_{nj}^2 \end{pmatrix}$$

или

Если $Q'' < Q'$

$$\Sigma_0 = \begin{pmatrix} \frac{1}{x_{1j}^2} & 0 & 0 & 0 \\ 0 & \frac{1}{x_{2j}^2} & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \frac{1}{x_{nj}^2} \end{pmatrix}$$

Тесты для проверки гетероскедастичности

Тест Глейзера

1. Будем предполагать, что $|e_i| = \alpha + \beta|x_{ii}|^\gamma + \delta_i$

2. Выдвигается гипотеза

$H_0 : \alpha = 0, \beta = 0$ - нет гетероскедастичности

3. Варьируя γ , оценивают уравнения регрессии. Если при оценивании значимым оказывается более одного уравнения, то выбирают уравнение с наибольшим коэффициентом детерминации

Тест Глейзера

4. В случае отклонения нулевой гипотезы структура матрицы Σ_0 имеет вид

$$\Sigma_0 = \begin{pmatrix} (\alpha + \beta |x_{1j}|^\gamma)^2 & 0 & 0 & 0 \\ 0 & (\alpha + \beta |x_{2j}|^\gamma)^2 & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & (\alpha + \beta |x_{nj}|^\gamma)^2 \end{pmatrix}$$

Тесты для проверки гетероскедастичности

Тест Парка (идея)

Парк предложил рассматривать следующую зависимость σ_i^2 от X_m

$$\sigma_i^2 = \sigma^2 \cdot (X_{m,i})^\beta \cdot e^{v_i} \quad (1)$$

где v_i удовлетворяют условиям 3-5 Гаусса-Маркова

В качестве оценки для σ_i^2 используются квадраты остатков исходной линейной модели e_i^2 и работают с моделью вида:

$$\ln(e_i^2) = \ln(\sigma^2) + \beta \cdot \ln(X_{m,i}) + v_i \quad (2)$$

Оценив параметры модели (2), проверяют, значим ли β . Если да, то отвергаем предположение об отсутствии гетероскедастичности.

Тест Парка (алгоритм)

Шаг 1. Находим МНК-оценки параметров исходной ЛММР.

Вычисляем остатки e_i

Шаг 2. Находим МНК-оценки параметров регрессии $\ln(e_i^2) = \ln(\sigma^2) + \beta \cdot \ln(X_{m,i}) + v_i$ по всем X_1, \dots, X_k .

Шаг 3. Проверяем значимость коэффициентов

Шаг 4. Делаем выводы об отсутствии или наличии гетероскедастичности

Шаг 4а. Формируем оценку матрицы Σ_0 (при необходимости)

Тест Бройша-Пагана-Годфри (идея)

Идея: дисперсии могут зависеть от абсолютно любой переменной, группы переменных или функционально преобразованных переменных.

Предполагается, что дисперсии ошибок описываются следующей функцией:

$$\sigma_i^2 = f(\theta_0 + \theta_1 Z_{1,i} + \dots + \theta_m Z_{m,i}),$$

где Z — любой ряд нестохастических данных (некоторыми Z могут быть Y , Y^2).

В частном случае можно рассматривать линейную зависимость дисперсий от Z вида $\sigma_i^2 = \theta_0 + \theta_1 Z_{1,i} + \dots + \theta_m Z_{m,i}$ и если $\theta_1 = \dots = \theta_m = 0$, то $\sigma_i^2 = \theta_0$ (то есть константе), и, следовательно, остатки гомоскедастичны.

Тест Бройша-Пагана-Годфри (алгоритм)

Шаг 1. Находим МНК-оценки параметров исходной ЛММР. Вычисляем остатки e_i .

Шаг 2. Рассчитываем оценку дисперсии по методу максимального правдоподобия

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n}.$$

Шаг 3. Строим ряд величин $p_i = \frac{e_i^2}{\tilde{\sigma}^2}$

Шаг 4. Находим МНК-оценки регрессии $p_i = \theta_0 + \theta_1 Z_{1,i} + \dots + \theta_m Z_{m,i} + v_i$.

Рассчитываем Q_{fact}

Шаг 5. В условиях справедливости нулевой гипотезы и при нормальности e_i статистика $\chi^2 = \frac{1}{2} \cdot Q_{fact}$ распределена по закону хи-квадрат с m степенями свободы.

Шаг 6 Делаем выводы об отсутствии или наличии гетероскедастичности

Шаг 6а. Формируем оценку матрицы Σ_0 (при необходимости)

Общий тест гетероскедастичности Уайта (алгоритм)

Шаг 1. Находим МНК-оценки параметров исходной ЛММР.

Вычисляем остатки e_i .

Шаг 2. Рассчитываем МНК-оценку уравнения регрессии по всем факторным признакам, их квадратам и всем попарным произведениям

Шаг 3. Рассчитываем \hat{R}^2 .

Шаг 4. При гомоскедастичности (все коэффициенты, кроме свободного члена, незначимы) статистика $n \cdot \hat{R}^2$ распределена по закону хи-квадрат с количеством степеней свободы, равным числу оцениваемых параметров в модели за исключением свободного члена.

Шаг 5 Делаем выводы об отсутствии или наличии гетероскедастичности

Шаг 5а. Формируем оценку матрицы $\hat{\Sigma}_0$ (при необходимости)

1. Стандартные ошибки в форме Уайта.

Предположим, что матрица ковариаций регрессионных остатков – диагональная. Тогда поскольку $b_{\text{МНК}} = \beta + (X^T X)^{-1} X^T \varepsilon$, то

$$\hat{\Sigma}_b = M\left[(X^T X)^{-1} X^T \varepsilon \varepsilon^T X (X^T X)^{-1}\right] = (X^T X)^{-1} X^T \hat{\Sigma}_\varepsilon X (X^T X)^{-1} = n(X^T X)^{-1} \left(\frac{1}{n} X^T \Sigma_\varepsilon X\right) (X^T X)^{-1}$$

Рассмотрим матрицу $X^T \Sigma_\varepsilon X$. Имеем $(X^T \Sigma_\varepsilon X)_{ij} = \sum_{s=1}^n x_{si} \sigma_s^2 x_{sj}$. Обозначим

через $X_s^T (1 \times k)$ - векторы-строки X . Тогда $X^T \Sigma_\varepsilon X = \sum_{s=1}^n \sigma_s^2 X_s X_s^T$. Уайт

показал, что

$\hat{\Sigma}_b = n(X^T X)^{-1} \left(\frac{1}{n} \sum_{s=1}^n \sigma_s^2 X_s X_s^T\right) (X^T X)^{-1}$ является состоятельной оценкой

ковариационной матрицы коэффициентов регрессии. Стандартные ошибки, рассчитанные по данной формуле называются стандартными ошибками в форме Уайта.

2. Стандартные ошибки в форме Невье-Веста.

Для более сложного случая, когда в ковариационной матрице регрессионных остатков неизвестные элементы стоят не только на главной диагонали, но и на соседних диагоналях, отстоящих от главной не более, чем на L (т.е. $\omega_{ij} = 0, |i - j| > L$). Невье и Вест показали, что оценка

$$\hat{\Sigma}_b = n(X^T X)^{-1} \left(\frac{1}{n} e_s^2 x_s x_s^T + \frac{1}{n} \sum_{j=1}^L \sum_{t=j+1}^n \omega_j e_t e_{t-j} (x_t x_{t-j}^T + x_{t-j} x_t) \right) (X^T X)^{-1} \quad \text{является}$$

состоятельной оценкой ковариационной матрицы коэффициентов регрессии.

Существует несколько способов выбора весовых коэффициентов ω_j :

- наиболее простым $\omega_j = 1$. Однако при таком выборе матрица может оказаться неположительно-определенной.

- $\omega_j = 1 - \frac{j}{L+1}$ (Бартлетт).

$$1 - 6\left(\frac{j}{L+1}\right)^2 + 6\left(\frac{j}{L+1}\right)^3, 1 \leq j \leq \frac{L+1}{2}$$

- $\omega_j = 2\left(1 - \frac{j}{L+1}\right)^2, \frac{L+1}{2} < j \leq L$ (Парзен).