Обобщенная линейная модель множественной регрессии с гетероскедастичными остатками

Линейная модель множественной регрессии

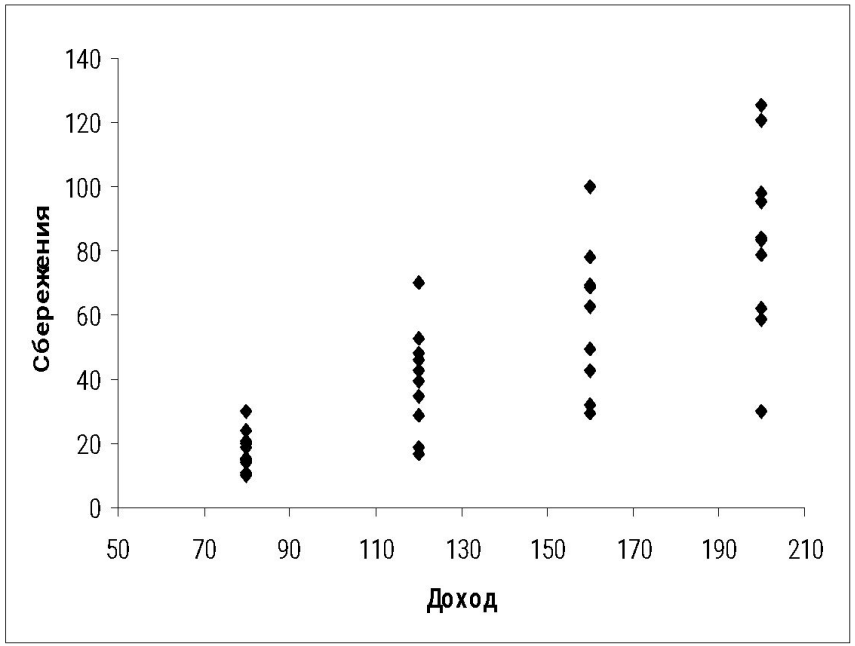
$$\overline{Y} = X\overline{\beta} + \overline{\varepsilon},$$

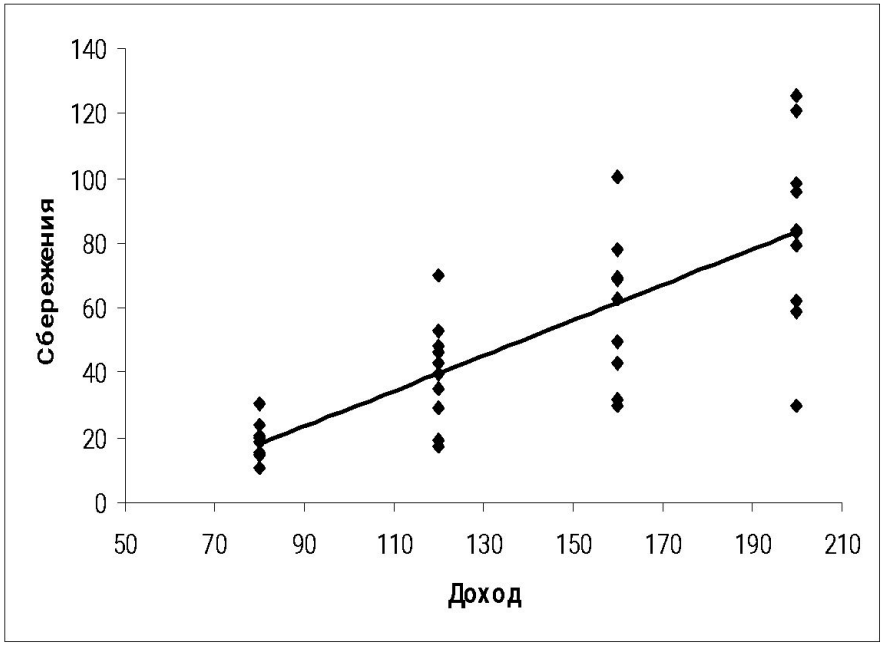
для которой нарушено 4 условие Гаусса-Маркова называется **обобщенной линейной моделью множественной регрессии (ОЛММР) с гетероскедастичными остатками,** а именно:

1) $x_1, \ldots, x_k$ – 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) $\operatorname{cov}(\varepsilon_i, \varepsilon_j) = M(\varepsilon_i \cdot \varepsilon_j) = 0,\ i \neq j,\ i = \overline{1,n}\ j = \overline{1,n}$

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

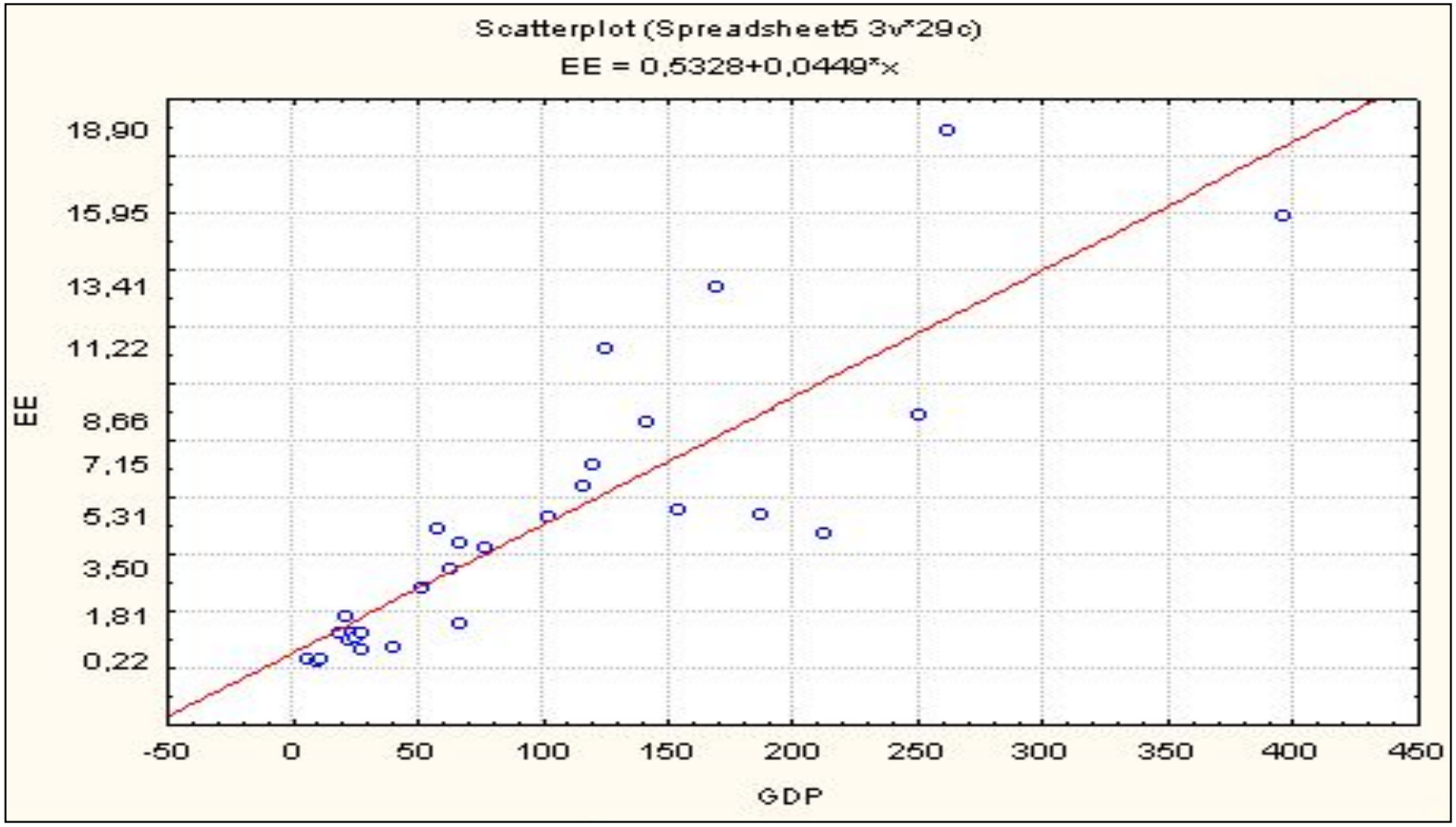
Пример 1. При исследовании среднедушевых сбережений и дохода в семьях разброс в данных будет выше для семей с более высокими доходами. Это означает, что дисперсия зависимых величин, а, следовательно, и случайных ошибок не постоянны.





Слайд №3

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



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 *di* — разность между рангами значений *xi* и *ei*

2. Выдвигается гипотеза

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

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

3. Гипотезе проверяется на основе статистика

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

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

$H_0 : \sigma_1^2 = \sigma_2^2 = \boxtimes = \sigma_n^2$ (нет гетероскедастичности)

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

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

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

Тест Голдфелда—Квандта

5. Строится статистики $F = \frac{\max(Q';Q'')}{\min(Q';Q'')}$

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

Если Q2>Q1

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

или

Если Q2<Q1

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

Тесты для проверки гетероскедастичности
Тест Глейзера

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

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

$$\hat{\Sigma}_0 = \begin{pmatrix} (\alpha + \beta \mid x_{1j} \mid^{\gamma})^2 & 0 & 0 & 0 \\ 0 & (\alpha + \beta \mid x_{2j} \mid^{\gamma})^2 & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & (\alpha + \beta \mid x_{nj} \mid^{\gamma})^2 \end{pmatrix}$$

Тесты для проверки гетероскедастичности
Тест Парка (идея)

Парк предложил рассматривать следующую зависимость σ_i^2 от X_m

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

где ν_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:

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

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 $X1, \ldots, Xk$.

Шаг 3. Проверяем значимость коэффициентов

Шаг 4. Делаем выводы об отсутствии или наличии гетероскедастичности

Шаг 4а. Формируем оценку матрицы $\hat{\Sigma}_0$ (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:

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

где Z – любой ряд нестохастических данных (некоторыми Z могут быть $\hat{Y}$, $\hat{Y}^2$).

В частном случае можно рассматривать линейную зависимость дисперсий от Z вида $\hat{\sigma}_i^2 = \theta_0 + \theta_1 Z_{1,i} + ... + \theta_m Z_{m,i}$ и если $\theta_1 = ... = \theta_m = 0$, то $\hat{\sigma}_i^2 = \theta_0$ (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 $\tilde{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i}{n}$.

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

Шаг 4. Находим МНК-оценки регрессии $p_i = \theta_0 + \theta_1 Z_{1,i} + \ldots + \theta_m Z_{m,i} + v_i$. 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 $\hat{\Sigma}_0$ (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 $\overline{R^2}$.

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

Шаг 5 Делаем выводы об отсутствии или наличии гетероскедастичности

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

Слайд №10

1. Стандартные ошибки в форме Уайта.

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

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

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

$\hat{\Sigma}_b = n(X^T X)^{-1}(\frac{1}{n} e_s^2 x_s x_s^T)(X^T X)^{-1}$ является состоятельной оценкой ковариационной матрицы коэффициентов регрессии. Стандартные ошибки, рассчитанные по данной формуле называются стандартными ошибками в форме Уайта.

Слайд №11

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

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

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).
- $\omega_j = \begin{cases} 1 - 6(\frac{j}{L+1})^2 + 6\left(\frac{j}{L+1}\right)^3, 1 \le j \le \frac{L+1}{2} \\ 2(1 - \frac{j}{L+1})^2, \frac{L+1}{2} < j \le L \end{cases}$ (Парзен).