Лекция № 3. Предпосылки метода наименьших квадратов. Обобщенный МНК

Вопросы

- 1. Предпосылки МНК и способы проверки их выполнения.
- 2. Свойства оценок, полученных с помощью МНК.
- 3. Обобщенный МНК.

1. При оценке параметров уравнения регрессии с помощью МНК делаются определенные предпосылки относительно случайной составляющей ε.

В модели

$$y = a + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p + \varepsilon$$

случайная составляющая *ε* представляет собой ненаблюдаемую величину.

После получения оценок параметров модели можно получить оценки *ε*, вычисляя разности фактических и теоретических значений результативного признака *y*. Так как они не являются реальными случайными остатками, их можно считать некоторой выборочной реализацией неизвестного остатка заданного уравнения, т.е. ε_i .

При изменении спецификации модели, добавлении в нее новых наблюдений выборочные остатки ε_i могут меняться. Поэтому в задачу регрессионного анализа входит не только построение самой модели, но и исследование случайных отклонений ε_i, т.е. остаточных величин.

Проверяя статистическую достоверность коэффициентов регрессии и корреляции, мы останавливались на t-критерии Стьюдента, *F*-критерии Фишера. При этом делались предположения относительно поведения остатков ε_i -

это независимые случайные величины; их среднее значение равно 0; они имеют постоянную дисперсию и подчиняются нормальному закону распределения. **Эти предположения являются** ***условиями теоремы Гаусса-Маркова.***

2. Статистические проверки параметров регрессии, показателей корреляции основаны на непроверяемых предпосылках распределения случайной составляющей ε_i. Они носят лишь предварительный характер. Уже после построения уравнения регрессии проводится проверка наличия у оценок ε_i тех свойств, которые изначально предполагались.

- **Речь идет о том, что оценки параметров регрессии должны быть *несмещенными*, *состоятельными* и *эффективными*. Эти свойства оценок, полученных по МНК, имеют чрезвычайно важное практическое значение в использовании результатов регрессии и корреляции.**

- **Напомним, что *несмещенность* оценки означает, что ее математическое ожидание равно оцениваемому параметру, а математическое ожидание остатков равно нулю. Следовательно, при большом числе выборочных оцениваний остатки не будут накапливаться и найденный параметр регрессии b_i**

можно рассматривать как среднее значение из возможного большого количества несмещенных оценок. Несмещенные оценки можно сравнивать по разным исследованиям.

Эффективность **оценок означает, что они характеризуются наименьшей дисперсией. В практических исследованиях это означает возможность перехода от точечного оценивания к интервальному.**

Степень реалистичности доверительных интервалов параметров регрессии обеспечивается, если оценки будут не только несмещенными и эффективными, но и *состоятельными*. Состоятельность оценок характеризует увеличение их точности с увеличением объема выборки.

Большой практический интерес представляют те результаты регрессии, для которых доверительный интервал ожидаемого значения параметра регрессии b_i имеет предел значений вероятности, равный единице. То есть вероятность получения оценки на заданном расстоянии от истинного значения параметра близка к единице.

Указанные критерии оценок (несмещенность, состоятельность, эффективность) обязательно учитываются при разных способах оценивания.

МНК строит оценки регрессии на основе минимизации суммы квадратов остатков. Поэтому очень важно исследовать их поведение.

Условия, необходимые для получения несмещенных, состоятельных и эффективных оценок, представляют собой *предпосылки МНК*, соблюдение которых желательно для получения достоверных результатов регрессии.

Исследования остатков ε_i предполагают проверку наличия следующих *пяти предпосылок МНК*:

1) случайный характер остатков;

2) нулевая средняя величина остатков, не зависящая от x_i;

3) *гомоскедастичность* – дисперсия каждого отклонения ε_i одинакова для всех значений x;

4) отсутствие автокорреляции остатков. Значения остатков ε_i распределены независимо друг от друга;

5) остатки подчиняются нормальному распределению.

Если хотя бы одна предпосылка не выполняется, следует корректиро-вать модель.

- **Для проверки *первой предпосылки* строится график зависимости остатков ε_i от теоретических значений результативного признака .**

  Если все значения остатков ε_i размещаются в горизонтальной полосе, то остатки представляют собой случайные величины и МНК оправдан, теоретические значения $\hat{y}_x$ хорошо аппроксимируют фактические значения *y* (рис. 1).

*

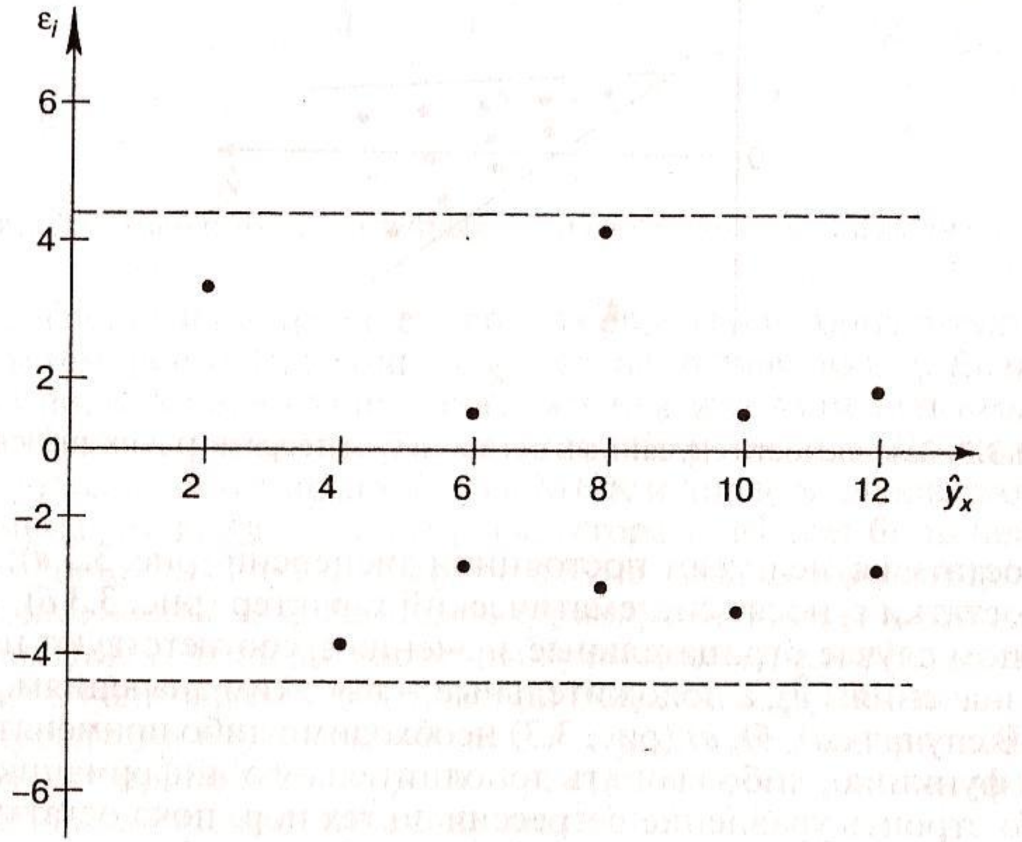


Рис.1. Зависимость случайных остатков $\boldsymbol{\varepsilon_i}$ от теоретических значений $\boldsymbol{\hat{y}_x}$

- **Если же зависимость остатков ε_i от $\hat{y}_x$ проявляется в том, что:**

  а) остатки ε_i не случайны;

  б) остатки не имеют постоянной дисперсии;

  в) остатки носят систематический характер, то нужно либо применять другую функцию, либо вводить дополнительную информацию и заново строить уравнение регрессии до тех пор, пока остатки ε_i не будут случайными величинами.

*

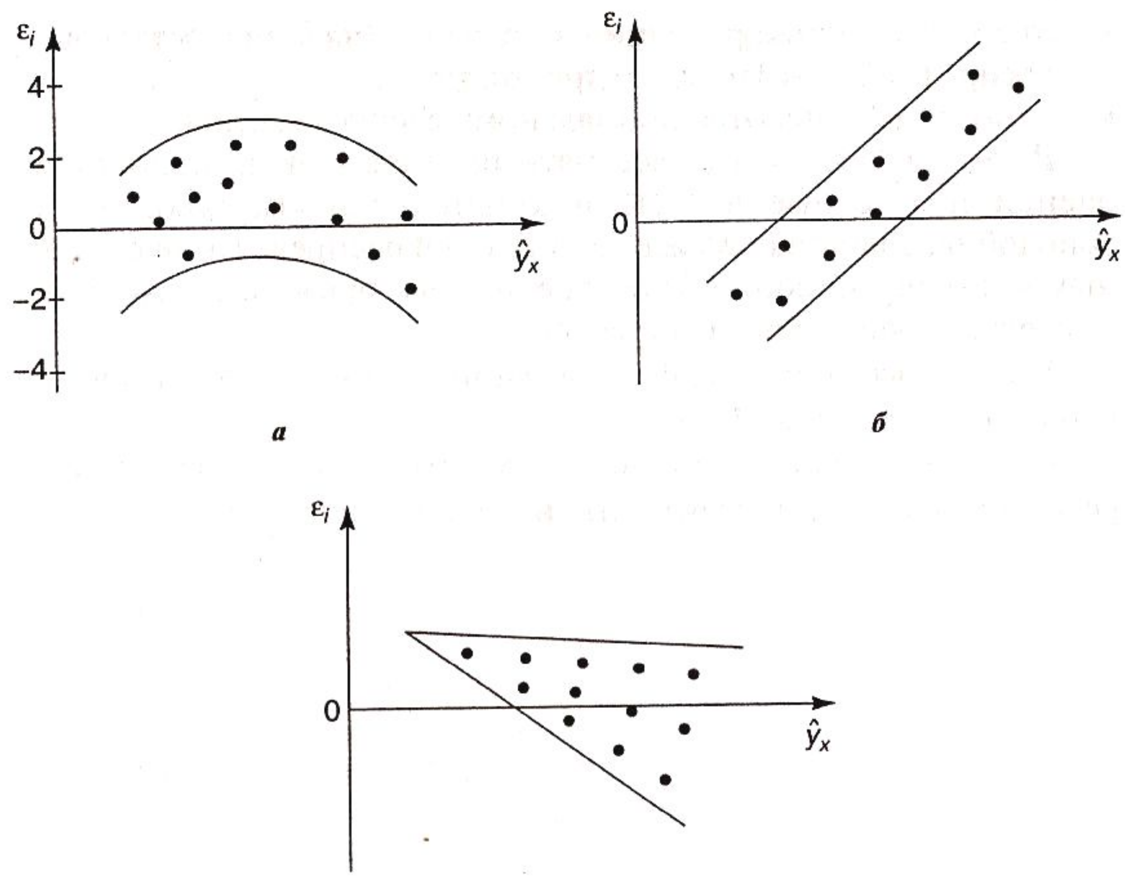


Рис. 2. Зависимость случайных остатков $\boldsymbol{\varepsilon_i}$ от теоретических значений $\boldsymbol{\hat{y}_x}$

- ***Вторая предпосылка*** **МНК относительно нулевой средней величины остатков означает, что**

$$\sum\left(y-\hat{y}_x\right)=0.$$

Это выполнимо для линейных моделей и моделей, нелинейных относительно включаемых переменных.

А для моделей, нелинейных относительно оцениваемых параметров и приводимых к линейному виду с помощью логарифмирования, средняя ошибка равна нулю для логарифмов исходных данных.

Так, для модели вида

$$y = ax_1^{b_1} \cdot x_2^{b_2} \cdot \ldots \cdot x_p^{b_p} \cdot \varepsilon \quad имеем, \quad \sum \left(\ln y - \ln \hat{y}_x \right) = 0 .$$

Кроме того, несмещенность оценок коэффициентов регрессии, полученных МНК, зависит также от независимости случайных остатков от величин *x*, что также исследуется в рамках соблюдения второй предпосылки МНК. С этой целью строится график зависимости случайных остатков *ε* от факторов x_i, включенных в регрессию.

- **Если остатки на графике расположены в виде горизонтальной полосы, то они независимы от значений x_j. Если же график показывает наличие указанной зависимости, то модель неадекватна (рис. 2).**

* *

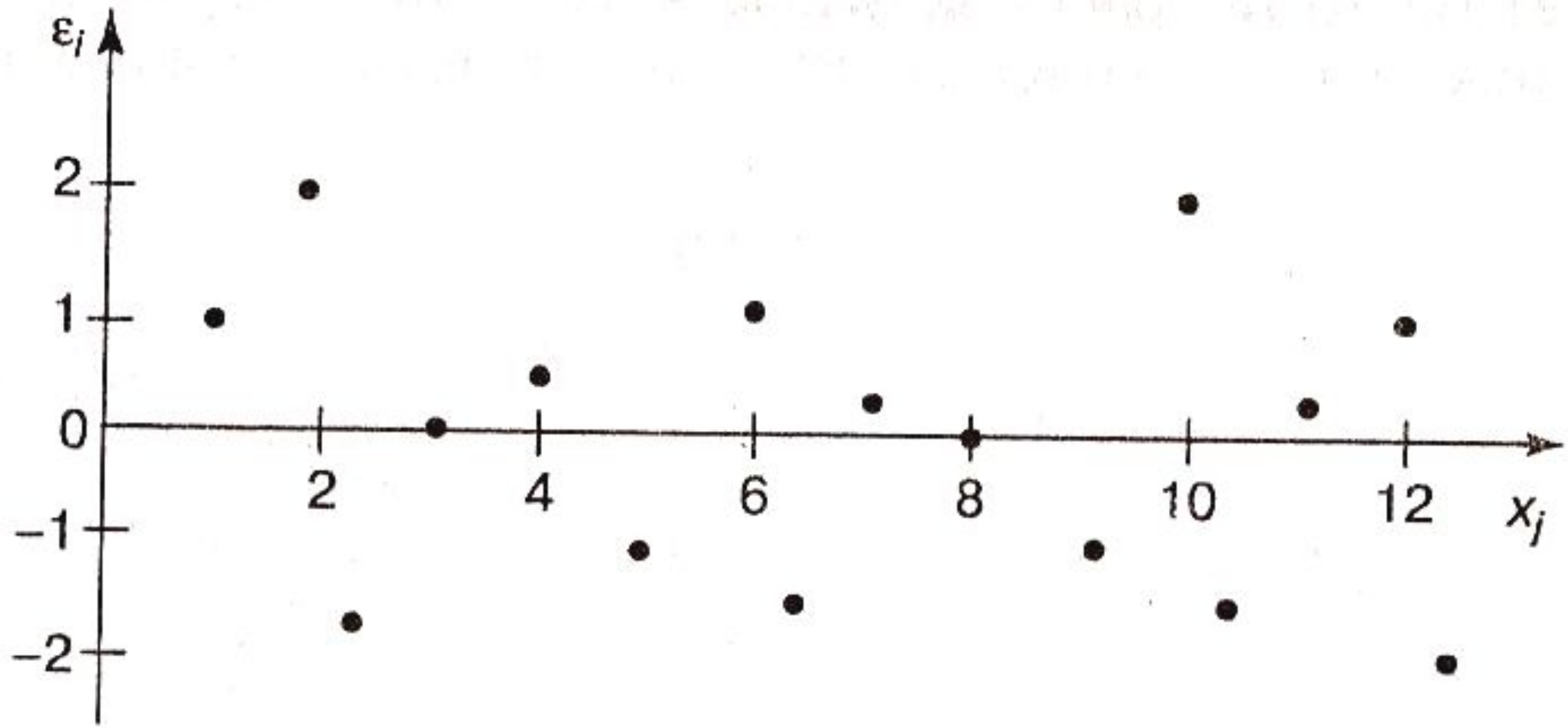


Рис. 3. Зависимость случайных остатков $\boldsymbol{\varepsilon_i}$ от величины фактора $\boldsymbol{x_j}$.

Причины **неадекватности могут быть разные: 1) нарушение третьей предпосылки МНК (дисперсия остатков не постоянна для каждого значения фактора x_j);**

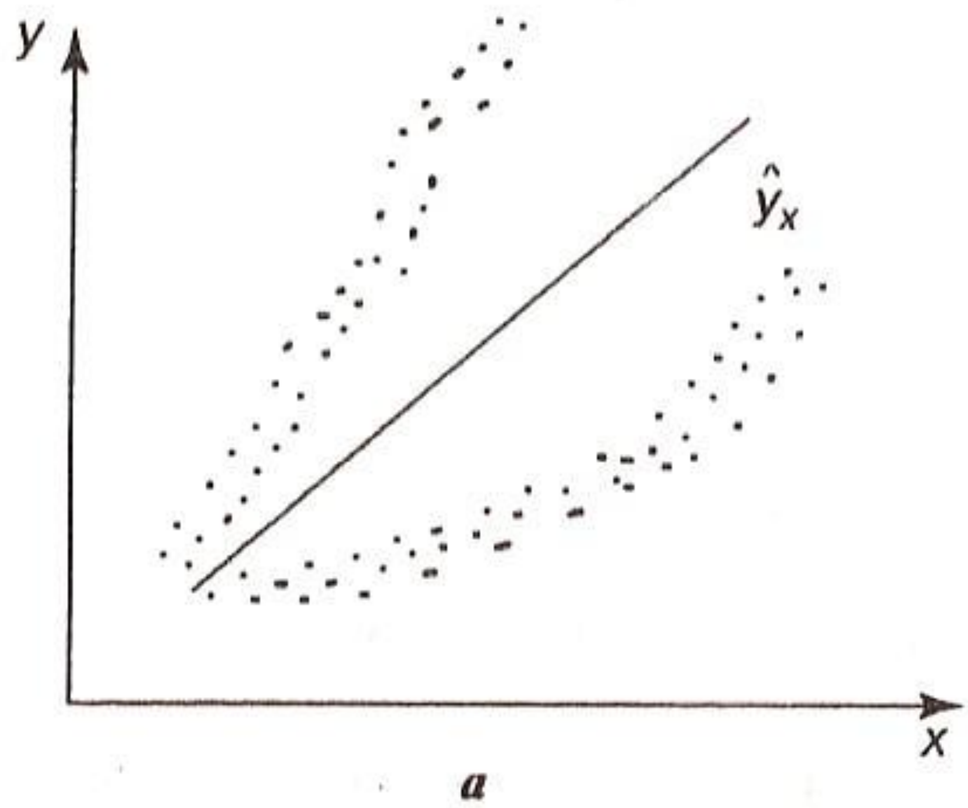
2) неправильная спецификация модели, и в нее необходимо ввести дополнительные члены от x_j, например, x_j^2, или преобразовать значения y. Скопление точек в определенных участках значений фактора x_j говорит о наличии систематической погрешности модели.

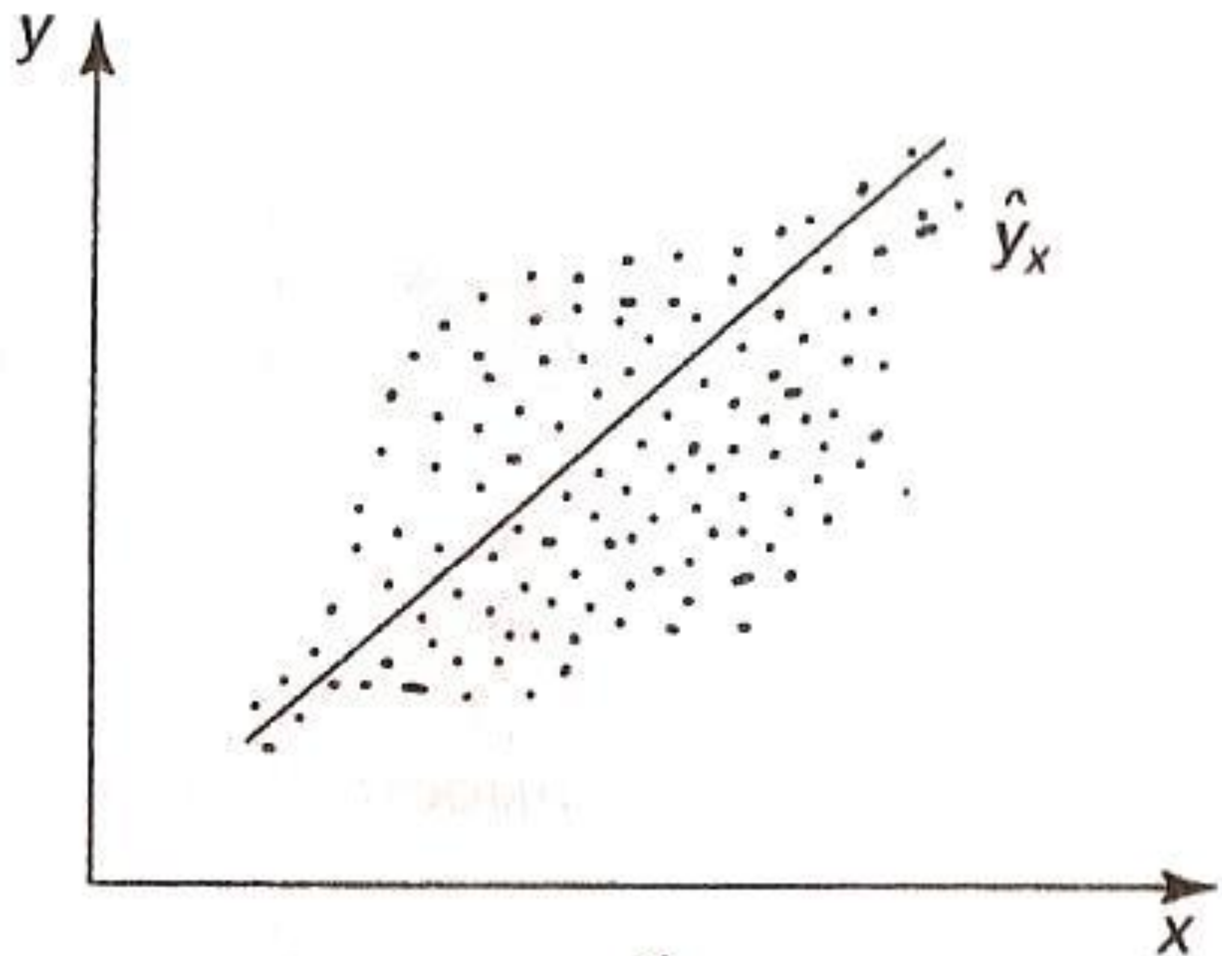
Предпосылка о нормальном распределении остатков позволяет проводить проверку параметров регрессии и корреляции с помощью критериев *t*, *F*. Вместе с тем оценки регрессии, найденные с применением МНК, обладают хорошими свойствами даже при отсутствии нормального распределения остатков, т.е. при нарушении пятой предпосылки МНК.

Для получения состоятельных оценок параметров регрессии по МНК совершенно необходимо соблюдение третьей и четвертой предпосылок.

В соответствии с *третьей предпосылкой МНК* требуется, чтобы дисперсия остатков была *гомоскедастичной*. Это значит, что для каждого значения фактора x_j остатки ε_i имеют одинаковую дисперсию. В противном случае имеем *гетероскедастичность*.

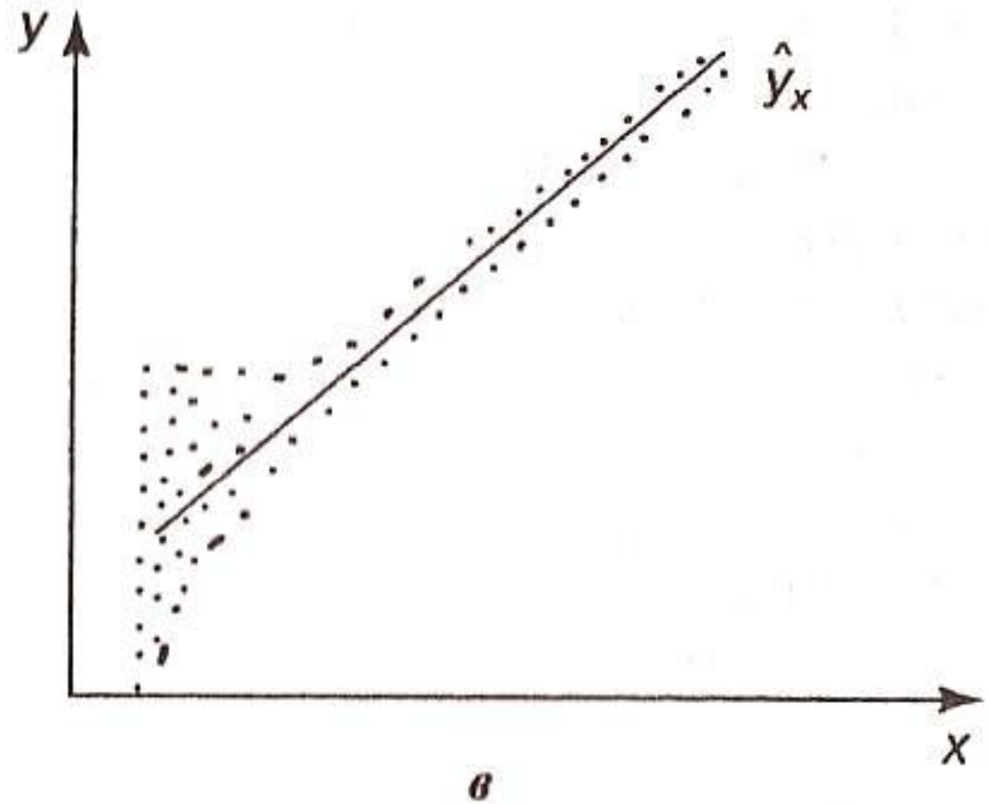
Наличие гетероскедастичности можно наглядно видеть из поля корреляции (рис. 4).





б

- *



- ***Рис. 4.*** Примеры гетероскедастичности:

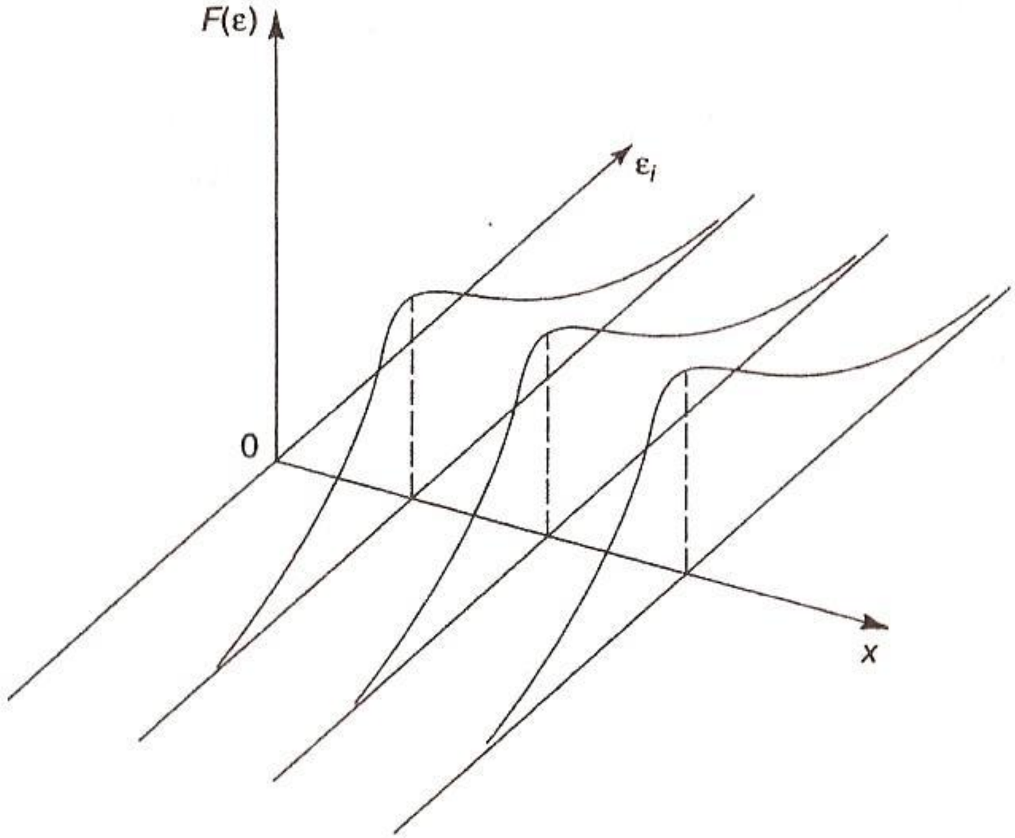
а) дисперсия остатков растет по мере увеличения x;

б) дисперсия остатков достигает максимальной величины при средних значениях переменной *x* и уменьшается при минимальных и максимальных значениях *x*;

в) максимальная дисперсия остатков при малых значениях *x* и дисперсия остатков однородна по мере увеличения значений *x.*

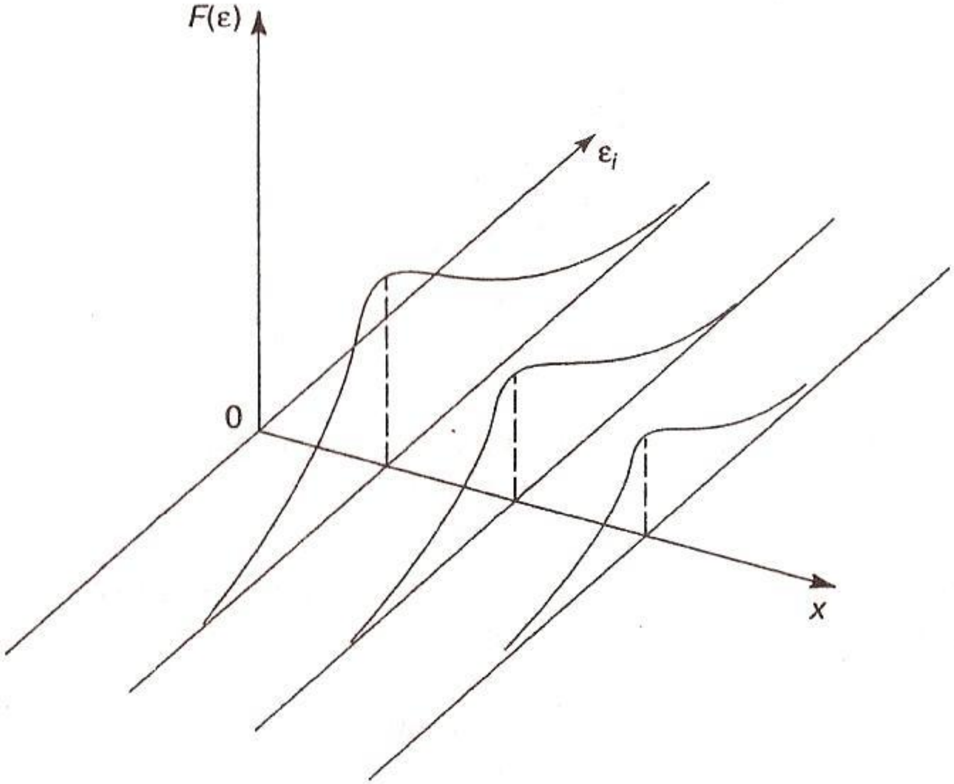
В случае гомоскедастичности для каждого значения x_i распределения остатков одинаковы, а в случае гетероскедастичности при переходе от одного значения x_i к другому меняется диапазон варьирования остатков.

- *



- ***Рис. 5.*** Гомоскедастичность остатков

- *



- ***Рис. 6.*** Гетероскедастичность остатков

Наличие гомоскедастичности или гетероскедастичности можно видеть и по рассмотренному выше графику зависимости остатков ε_i от теоретических значений результативного признака $\hat{y}_x$. Так, для рисунка 4а) зависимость остатков от $\hat{y}_x$ представлена на рис. 7.

*

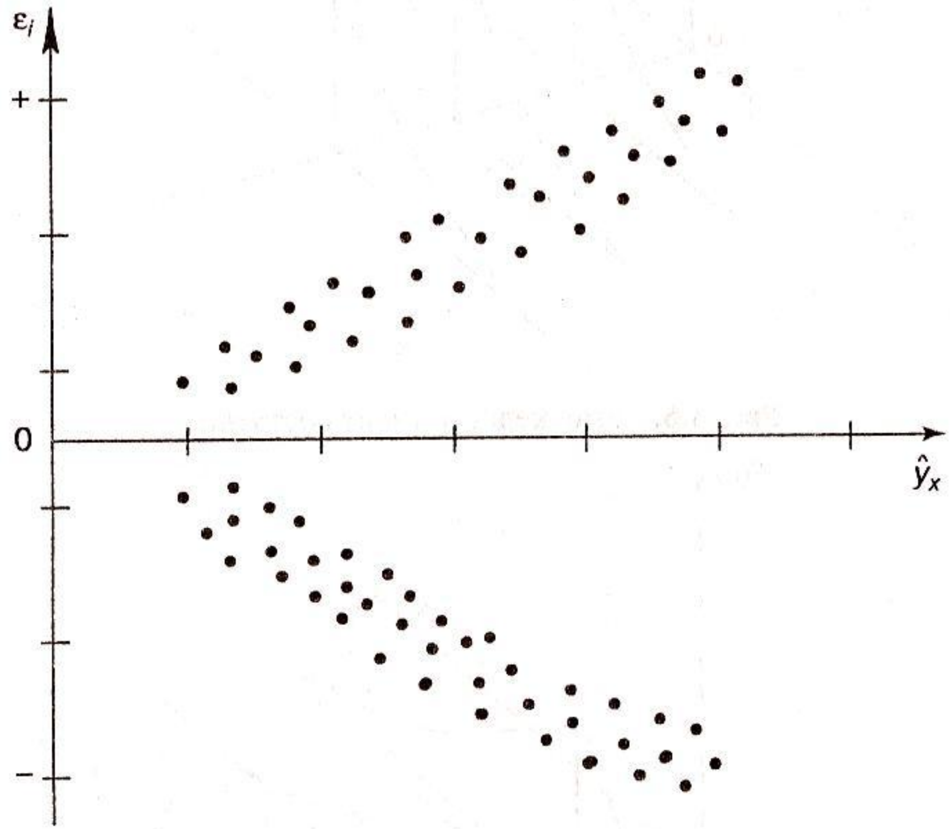
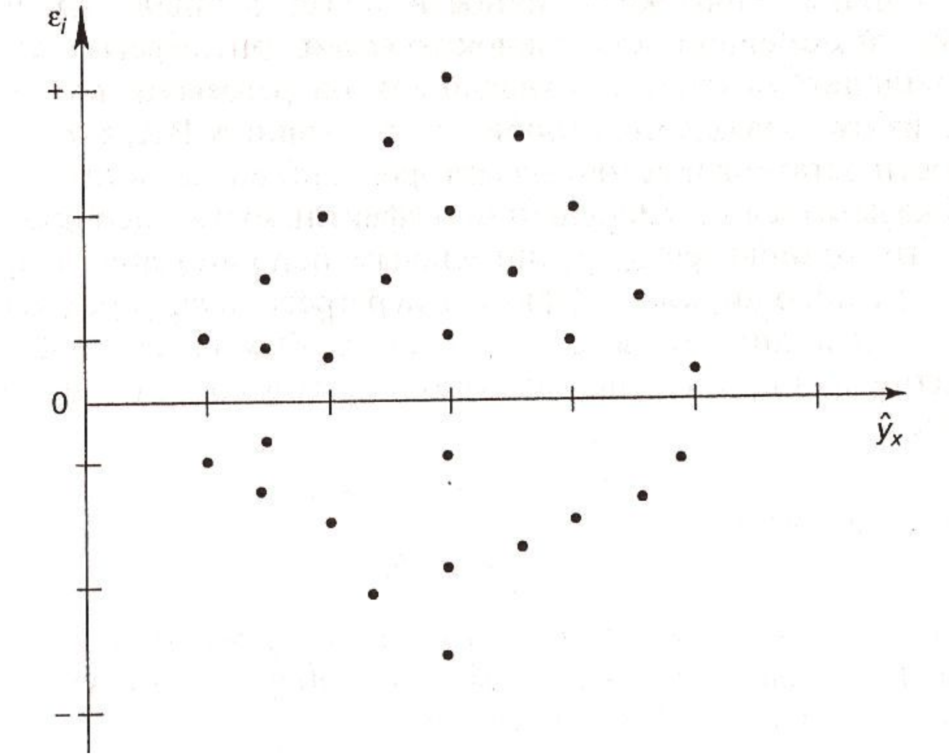


Рис. 7. Гетероскедастичность: большая дисперсия ε_i для больших значений $\hat{y}_x$.

Соответственно для зависимостей, изображенных на полях корреляции рис. 4 б) и в), гетероскедастичность остатков представлена на рис. 8 и 9.

*



- ***Рис. 8***. Гетероскедастичность, соответствующая полю корреляции рис. 4б)

*

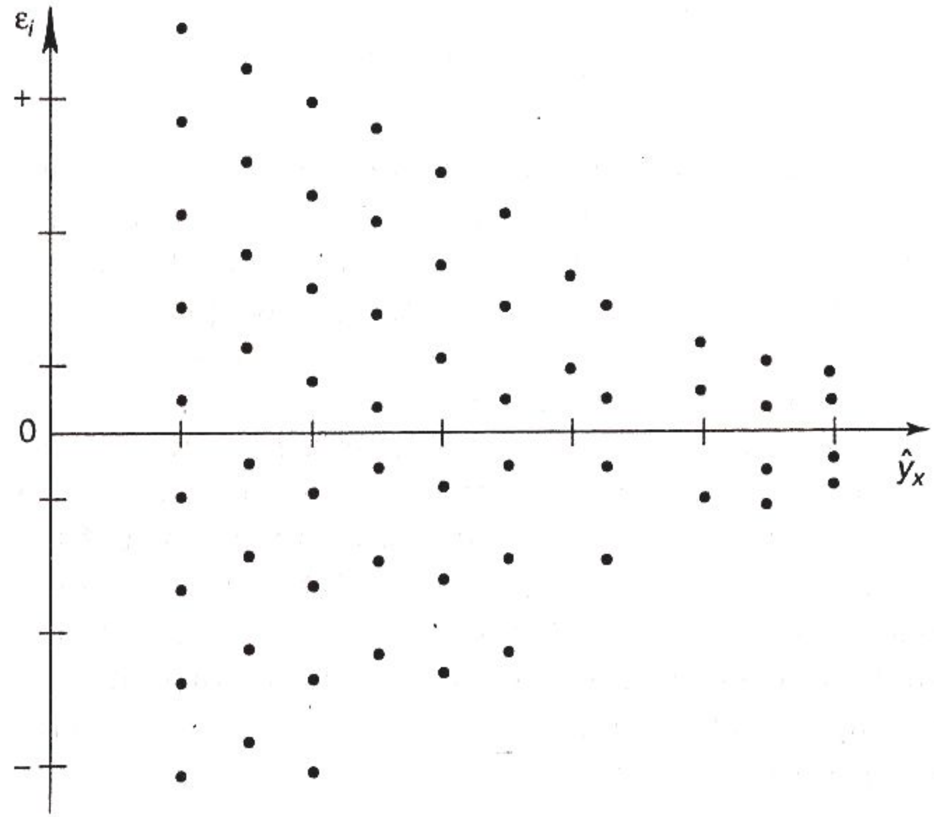


Рис. 9. Гетероскедастичность, соответствующая полю корреляции рис. 4в)

Наличие гетероскедастичности может в отдельных случаях привести к смещенности оценок коэффициентов регрессии, хотя несмещенность этих оценок зависит в основном от соблюдения второй предпосылки МНК. Гетероскедастичность будет сказываться на уменьшении эффективности оценок b_i.

Практически при нарушении гомоскедастичности мы имеем неравенства:

$$\sigma^2_{\varepsilon_i} \neq \sigma^2_{\varepsilon_j} \neq \sigma^2 , \quad i \neq j ,$$

и можно записать

$$\sigma^2_{\varepsilon_i} = \sigma^2 \cdot K_i \ .$$

При этом величина K_i может меняться при переходе от одного значения фактора x_i к другому.

Это означает, что сумма квадратов отклонений для зависимости

$$\hat{y}_x = a + bx$$

при наличии гетероскедастичности должна иметь вид:

$$S_{гетеро} = \sum \frac{1}{K_i}\left(\hat{y}_i - a - bx_i\right)^2 .$$

При минимизации этой суммы квадратов отдельные ее слагаемые взвешиваются: наблюдениям с наибольшей дисперсией придается пропорционально меньший вес. Иными словами, для учета систематического влияния неоднородных элементов K_i вклад каждой пары x_i с y_i в сумму квадратов остатков должен быть дисконтирован.

Задача состоит в том, чтобы определить величину K_i и внести поправку в исходные переменные.

С этой целью рекомендуется использовать *обобщенный метод наименьших квадратов*, который эквивалентен обыкновенному МНК, примененному к преобразованным данным.

3. *Обобщенный МНК* применяется при нарушении гомоскедастичности и наличии автокорреляции ошибок.

ОМНК применяется к преобразованным данным и позволяет получать оценки, обладающие не только свойством несмещенности, но и имеющие наименьшие выборочные дисперсии. Остановимся на использовании ОМНК для корректировки гетероскедастичности.

Как и раньше, будем предполагать, что среднее значение остатков равно нулю, а дисперсия не остается постоянной для разных значений фактора, а изменяется пропорционально величине K_i , т.е.

$$\sigma_{\varepsilon_i}^2 = \sigma^2 \cdot K_i \ ,$$

где $\sigma_{\varepsilon_i}^2$ - дисперсия ошибки при конкретном i-м значении фактора;

σ^2 - постоянная дисперсия ошибки при соблюдении предпосылки о гомоскедастичности остатков;

K_i – коэффициент пропорциональности, меняющийся с изменением величины фактора, что и обусловливает неоднородность дисперсии.

В общем виде для уравнения

$$y_i = a + bx_i + \varepsilon_i \qquad при \qquad \sigma^2_{\varepsilon_i} = \sigma^2 \cdot K_i$$

модель примет вид:

$$y_i = \alpha + \beta x_i + \sqrt{K_i} \cdot \varepsilon_i \ .$$

В ней остаточные величины гетероскедастичны. Предполагая в них отсутствие автокорреляции, можно перейти к уравнению с гомоскедастичными остатками, поделив все переменные, зафиксированные в ходе i-го наблюдения, на $\sqrt{K_i}$.

Тогда дисперсия остатков будет величиной постоянной, т.е.

$$\sigma_{\varepsilon_i}^{2} = \sigma^2.$$

Таким образом, от регрессии *y* по *x* мы перейдем к регрессии на новых переменных:

$$\frac{y}{\sqrt{K}} \quad и \quad \frac{x}{\sqrt{K}}.$$

Уравнение регрессии примет вид:

$$\frac{y_i}{\sqrt{K_i}} = \frac{\alpha}{\sqrt{K_i}} + \beta \cdot \frac{x_i}{\sqrt{K_i}} + \varepsilon_i \ .$$

Исходные данные для данного уравнения будут иметь вид:

$$y = \begin{pmatrix} \dfrac{y_1}{\sqrt{K_1}} \\ \dfrac{y_2}{\sqrt{K_2}} \\ \dots \\ \dfrac{y_n}{\sqrt{K_n}} \end{pmatrix} \qquad x = \begin{pmatrix} \dfrac{x_1}{\sqrt{K_1}} \\ \dfrac{x_2}{\sqrt{K_2}} \\ \dots \\ \dfrac{x_n}{\sqrt{K_n}} \end{pmatrix}$$

По отношению к обычной регрессии уравнение с новыми, преобразованными, переменными представляет собой *взвешенную регрессию*, в которой переменные *x* и *y* взяты с весами

$$\frac{1}{\sqrt{K}} .$$

Оценка параметров нового уравнения с преобразованными переменными приводит к взвешенному методу наименьших квадратов, для которого необходимо минимизировать сумму квадратов отклонений вида

$$S = \sum \frac{1}{K_i} \cdot \left(y_i - a - bx_i\right)^2 .$$

Соответственно получим следующую систему нормальных уравнений:

$$\begin{cases} \sum \dfrac{y_i}{K_i} = a \cdot \sum \dfrac{1}{K_i} + b \cdot \sum \dfrac{x_i}{K_i}, \\ \sum \dfrac{y_i x_i}{K_i} = a \cdot \sum \dfrac{x_i}{K_i} + b \cdot \sum \dfrac{x_i^2}{K_i}. \end{cases}$$

Если преобразованные переменные *x* и *y* взять в отклонениях от средних уровней, то коэффициент регрессии *b* можно определить как

$$b = \frac{\sum \frac{1}{K} xy}{\sum \frac{1}{K} x^2}.$$

При обычном применении МНК для переменных в отклонениях от средних уровней коэффициент регрессии определяется по формуле

$$b = \frac{\sum xy}{\sum x^2}.$$

Таким образом, при использовании обобщенного МНК с целью корректировки гетероскедастичности коэффициент регрессии *b* представляет собой взвешенную величину по отношению к обычному МНК с весами *1/K*.

Рассмотрим данный подход для уравнения *множественной регрессии*.

Пусть рассматривается модель вида

$$y = a + b_1x_1 + b_2x_2 + \varepsilon,$$

для которой дисперсия остатков оказалась пропорциональной K^2_i, где K_i – коэффициент пропорциональности, принимающий различные значения для соответствующих i значений факторов x_1 и x_2.

Так как $\sigma_{\varepsilon_i}^2 = \sigma^2 \cdot K_i^2$,

рассматриваемая модель примет вид

$$y_i = a + b_1 x_{1i} + b_2 x_{2i} + K_i \varepsilon_i \ ,$$

где ошибки гетероскедастичны.

Для перехода к новому уравнению с гомоскедастичными остатками разделим все члены исходного уравнения на коэффициент пропорциональности K.

Тогда

$$\frac{y_i}{K_i} = \frac{a}{K_i} + b_1 \frac{x_{1i}}{K_i} + b_2 \frac{x_{2i}}{K_i} + \varepsilon_i \ .$$

- **Это уравнение не содержит свободного члена. Вместе с тем, найдя переменные в новом преобразованном виде и применяя к ним обычный МНК, получим иную спецификацию модели:**

$$\frac{y_i}{K_i} = A + b_1 \frac{x_{1i}}{K_i} + b_2 \frac{x_{2i}}{K_i} + \varepsilon_i \ .$$

Параметры такой модели зависят от концепции, принятой для коэффициентов пропорциональности K_i. В эконометрических исследованиях довольно часто выдвигается гипотеза, что остатки ε_i пропорциональны значениям фактора.

Так, если в уравнении

$$y = a + b_1 x_1 + b_2 x_2 + ... + b_p x_p + E$$

предположить, что $E = \varepsilon x_1$, т.е. $K = x_1$ и

$$\sigma^2_{\varepsilon i} = \sigma^2 x_1^2 \ ,$$

то ОМНК предполагает оценку параметров следующего трансформированного уравнения:

$$\frac{y}{x_1} = b_1 + b_2 \frac{x_2}{x_1} + ... + b_p \frac{x_p}{x_1} + \varepsilon \ .$$

Если предположить, что ошибки пропорциональны x_p, то модель примет вид:

$$\frac{y}{x_p} = b_p + b_1 \frac{x_1}{x_p} + \ldots + b_{p-1} \frac{x_{p-1}}{x_p} + \varepsilon .$$

Применение в этом случае обобщенного МНК приводит к тому, что наблюдения с меньшими значениями преобразованных переменных x/K имеют при определении параметров регрессии относительно больший вес, чем с первоначальными переменными.

Вместе с тем следует иметь в виду, что новые преобразованные переменные получают новое экономическое содержание и их регрессия имеет иной смысл, чем регрессия по исходным данным.