Эконометрика

Эконометрика и эконометрическое моделирование: основные понятия и определения

- Под *экономическим объектом* будем понимать любой элемент экономики (микроуровень: фирмы, семьи, предприятия; мезоуровень: регионы, отдельный сектор экономики, отрасли, корпорации; макроуровень: экономика страны в целом).
- *Экономические переменные* — это набор количественных характеристик, описывающий деятельность экономического объекта.

- *Эконометрика* — это научная дисциплина, объединяющая совокупность теоретических результатов, приемов, методов и моделей, предназначенная для того, чтобы на базе: экономической теории, экономической статистики и экономических измерений, математико-статистического инструментария придавать конкретное ко личественное выражение общим (качественным) закономерностям, обусловленным экономической теорией).

Эконометрика является одним из разделов математического моделирования экономических процессов, который базируется:

- на экономической теории;
- экономической статистике и экономических измерениях;
- математико-статистическом инструментарии,

и предназначена для построения *эконометрических моделей,* которые ис пользуются для оценивания и прогнозирования значений экономических переменных, недоступных для измерения.

Этапы построения эконометрических моделей и принципы спецификации

Построение эконометрических моделей (как и экономико-математических) выполняется в несколько этапов:

1) спецификация модели;
2) сбор статистической информации об объекте исследования;
3) оценка параметров модели (параметризация, настройка);
4) проверка адекватности модели (верификация).

- Экономико-математическая модель объекта (математическая модель экономического объекта) представляет собой *математически выраженную* связь между его экономическими переменными.
- Это либо набор графиков или таблиц, либо система математических уравнений (алгебраических, конечно-разностных, дифференциальных, интеральных и т. д.) и, возможно, неравенств, связывающих воедино экономические переменные объекта.

По отношению к выбранной спецификации все экономические переменные объекта подразделяются на два типа:

- *эндогенные*
- *экзогенные.*

Определение

- *Экзогенными* (независимыми) называются экономические переменные, значения которых определяются вне данной модели.

- *Эндогенными* (зависимыми) называются экономические переменные, значения которых определяются (объясняются) внутри модели в результате одновременного взаимодействия соотношений, образующих модель.

Определение

- При наличии хотя бы одной экзогенной переменной модель называется *открытой,* в противном случае — *замкнутой.*

Первый принцип спецификации

- *Экономико-математическая* модель строится по результатам математической формализации закономерностей общей экономической теории.

Второй принцип спецификации

- В правильно составленной спецификации содержится столько уравнений, сколько эндогенных переменных включается в модель

Третий принцип спецификации

- Учет фактора времени в экономических моделях, или датирование экономических переменных.

Определение

- Переменные модели называются *датированными,* если обозначена их зависимость от времени.

- Если экономические утверждения отражают *статическую* (относящуюся к одному периоду времени) взаимосвязь всех включённых в модель переменных, то значения таких переменных принято называть *пространственными данными.*

И надобности в их датировании нет.

- Если экономические утверждения отражают *динамическую* (зависящую от фактора времени) взаимосвязь включённых в модель переменных, то значения таких переменных датируют и называют *динамическими или временными рядами*

Определение

- *Лаговыми* называются экзогенные или эндогенные переменные экономической модели, датированные предыдущими моментами времени и находящиеся в уравнении с текущими переменными.

- Модели, включающие лаговые переменные, относятся к классу *динамических моделей.*

Определение

Предопределёнными называются лаговые и текущие экзогенные переменные, а также лаговые эндогенные переменные.

Четвертый принцип спецификации

- Включение случайных возмущений в спецификацию экономической модели.

- Экономические модели со случайными возмущениями принято называть *эконометрическими.*

- На первом этапе построения эконометрических моделей, то есть — спецификации модели привлекается общая экономическая теория и математика и не содержат информацию о конкретных значениях параметров модели.

- Поэтому для построения оценок (или прогнозов) значений эндогенных переменных необходимо привлечь результаты статистических наблюдений за данным экономическим объектом, полученные на *втором этапе построения модели.*

- Далее, на основании статистической информации при помощи статистических методов (как правило, методов регрессионного анализа) выполняется оценка параметров модели — *третий этап построения модели* (этап настройки).

- Таким образом, на втором и третьем этапах привлекается третья составляющая эконометрики — статистика (теория статистических измерений и математическая статистика).

- Следующий этап построения эконометрической модели — *верификация* (проверка адекватности модели).
- На данном этапе проверяется соответствие модели эмпирическим данным.

Структурная и приведенная формы эконометрических моделей

Для построения прогнозов эндогенных переменных необходимо выразить текущие эндогенные переменные модели в виде *явных функций* предопределённых переменных. Последняя спецификация, полученная путем включения случайных возмущений получена в результате математической формализации экономических закономерностей. Такая форма спецификации называется *структурной.* В общем случае в структурной спецификации эндогенные переменные не выражены в явном виде через предопределенные.

В модели равновесного рынка только переменная предложения выражена в явном виде через предопределенную переменную, поэтому для представления эндогенных переменных через предопределенные необходимо выполнить некоторые преобразования структурной формы. Решим систему уравнений для последний спецификации относительно эндогенных переменных.

Получим

$$Y_t^d = a_0 + a_1 p_t + a_2 x_t + u_t, \quad a_1 < 0, \quad a_2 > 0,$$

$$Y^s = b_0 + b_1 p_{t-1} + v_t, \quad b_1 > 0$$

$$Y_t^d = Y_t^s$$

Подставим первое и второе уравнения в третье,

$$a_0 + a_1 p_t + a_2 x_t + u_t = b_0 + b_1 p_{t-1} + v_t,$$

и выразим текущее значение цены равновесия через предопределенные

$$p_t = \frac{b_0 - a_0}{a_1} + \frac{b_1}{a_1} \cdot p_{t-1} - \frac{a_2}{a_1} \cdot x_t + \frac{v_t - u_t}{a_1} = c_0 + c_1 \cdot p_{t-1} + c_2 \cdot x_t + \varepsilon_t,$$

где $c_1 < 0$, $c_2 > 0$, $\varepsilon_t = \frac{v_t - a_0}{a_1}$.

Окончательно получим выражение спроса через предопределенные переменные:

$$Y_t^d = a_0 + a_1 \cdot \left(\frac{b_0 - a_0}{a_1} + \frac{b_1}{a_1} \cdot p_{t-1} - \frac{a_2}{a_1} \cdot x_t + \frac{v_t - u_t}{a_1} \right) + a_2 \cdot x_t + u_t = b_0 + b_1 \cdot p_{t-1} + v_t .$$

Таким образом, после преобразований спецификация модели принимает следующий вид:

$$Y_t^d = b_0 + b_1 p_{t-1} + v_t,$$

$$Y^s = b_0 + b_1 p_{t-1} + v_t,$$

$$p_t = c_0 + c_1 p_{t-1} + c_2 x_t + \varepsilon_t,$$

$$c_1 < 0, \quad c_2 > 0, \quad b_1 > 0$$

Таким образом, эндогенные переменные модели выражены в явном виде через предопределенные переменные. Такая форма спецификации получила название *приведенной.* В частном случае структурная и приведённая формы модели могут совпадать. При правильной спецификации модели переход от структурной к приведённой форме всегда возможен, обратный переход возможен не всегда.

Матричная запись структурной и приведенной форм моделей

Введем следующие обозначения:

- Y_t — вектор-столбец текущих значений эндогенных переменных;
- X_t — расширенный вектор-столбец предопределённых переменных, значения которых известны к моменту t;
- A и B — матрицы коэффициентов структурной формы модели (структурные коэффициенты);
- V_t — вектор-столбец текущих возмущений.

С учетом данных обозначений матричная запись структурной формы эконометрической модели принимает вид

$$A Y_t + B X_t = V_t .$$

Представим спецификацию модели равновесного рынка в матричной форме. Для этого предварительно в каждом уравнении системы перенесем все члены (кроме случайных возмущений) в левую часть:

$$Y_t^d - a_0 - a_1 p_t - a_2 x_t = u_t, \quad a_1<0, \quad a_2>0,$$
$$Y^s - b_0 - b_1 p_{t-1} = v_t, \quad b_1>0$$
$$Y_t^d - Y_t^s = 0$$

Элементы вектора эндогенных переменных следующие:

$$Y_t=(Y_t^d, Y_t^s, p_t)^T$$

Предопределенные переменные модели:

лаговое значение цены товара;

текущий доход потребителя.

Расширенный вектор X, включает три элемента:

$$X_t=(1,\ p_{t-1},\ x_t)^T$$

Элементами вектора V_t в модели являются текущие возмущения соответствующих поведенческих уравнений и нулевой элемент — правая часть уравнения тождества

$$V_t=(\ u_t,\ v_t\ ,\ 0)^T\ .$$

Матрицы структурных коэффициентов A и B состоят из следующих элементов:

$$A=\begin{pmatrix}1 & 0 & -a_1\\ 0 & 1 & 0\\ 1 & -1 & 0\end{pmatrix}$$

$$B=\begin{pmatrix}-a_0 & 0 & -a_2\\ -b_0 & -b_1 & 0\\ 0 & 0 & 0\end{pmatrix}$$

Матричное представление *приведённой формы* спецификации следующее:

$$Y_t = M X_t + U_t,$$

где *M*— матрица приведенных коэффициентов, то есть

$$M = -A^{-1} B.$$

Или

$$M = -A^{-1} \cdot B = \begin{pmatrix} 0 & -1 & -1 \\ 0 & -1 & 0 \\ \frac{1}{a_1} & -\frac{1}{a_1} & -\frac{1}{a_1} \end{pmatrix} \cdot \begin{pmatrix} -a_0 & 0 & -a_2 \\ -b_0 & -b_0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} b_0 & b_1 & 0 \\ b_0 & b_1 & 0 \\ \frac{b_0 - a_0}{a_1} & \frac{b_1}{a_1} & -\frac{a_2}{a_1} \end{pmatrix}$$

Окончательно получим

Окончательно получим

$$\begin{pmatrix} Y_t^d \\ Y_t^s \\ p_t \end{pmatrix} = \begin{pmatrix} b_0 & b_1 & 0 \\ b_0 & b_1 & 0 \\ \dfrac{b_0 - a_0}{a_1} & \dfrac{b_1}{a_1} & -\dfrac{a_2}{a_1} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ p_{t-1} \\ x_t \end{pmatrix} + \begin{pmatrix} v_t \\ v_t \\ u_t \end{pmatrix}.$$

Вектор случайных возмущений в приведённой форме получается в результате преобразования

$$U_t = A^{-1} V_t$$

Или в координатной форме

$$U_t = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 1 & 0 \\ -1/a_1 & 1/a_1 & 1/a_1 \end{pmatrix} \cdot \begin{pmatrix} u_t \\ v_t \\ 0 \end{pmatrix} = \begin{pmatrix} v_t \\ v_t \\ \dfrac{v_t - u_t}{a_1} \end{pmatrix}.$$

Пример. *Модель формирования национального дохода* (Дж. М. Кейнс)

Исследуемым экономическим объектом является закрытая национальная экономика без государственного вмешательства.

Экономические переменные модели:

$$Y, C, I,$$

где Y— уровень совокупного выпуска (национальный доход), C — объём потребления, I— величина инвестиций. I.

Требуется:

A. Составить спецификацию макромодели, позволяющей объяснять величины *Y* (национального дохода) и *C* (объем потребления) уровнем инвестиций

При составлении спецификации модели воспользоваться следующими утверждениями экономической теории:

1) потребление возрастает с увеличением совокупного выпуска, причём рост потребления происходит медленнее роста совокупного выпуска;
2) в закрытой экономике без государственного вмешательства потребление и инвестиции в сумме равны совокупному выпуску (тождество системы национальных

Б. Уточнить спецификацию путем датирования переменных. При да тировании экономических переменных данной модели следует учесть еще один факт экономической теории: текущее потребление зависит от сово купного выпуска предыдущего периода.

В. Уточнить спецификацию включением случайного возмущения.

Г. Составить приведенную форму спецификации.

Д. Записать структурную и приведенную формы в матричном виде.

Решение.

Воспользуемся *первым принципом* спецификации и формализуем экономические законы, характеризующие данный экономический объект.

А. Исходя из первой закономерности экономической теории, имеем:

$$C = a+bY, \quad 0<b<1, \quad a>0,$$

где *a* — базовое потребление, *b* — предельное потребление в зависимости от дохода (склонность к потреблению).

Из второй предпосылки следует тождество

$$Y = C + I.$$

Таким образом, *структурная форма модели,* полученная в результате формализации экономических закономерностей, имеет вид

$$C = a + bY, \quad 0<b<1, \quad a>0,$$

$$Y = C + I.$$

Вывод

Спецификация составлена правильно, так как в структурной форме, в соответствии со вторым принципом, должно быть два уравнения — модель включает две эндогенные переменные модели:

уровень дохода *Y;*

уровень потребления *C*.

Экзогенной переменной является *I* — уровень инвестиций.

Б. Третий принцип спецификации — датирование переменных. Необходимо уточнить спецификацию: датировать экономические переменные, т. е. учесть их зависимость от фактора времени.

При датировании экономических переменных данной модели следует учесть тот факт, что *текущее потребление зависит от совокупного выпуска предыдущего периода,* поэтому уточненная датированная структурная спецификация принимает вид:

$$C_t = a + bY_{t-1} \quad 0<b<1, \quad a>0,$$

$$Y_t = C_t + I_t.$$

В. Уточним спецификацию включением случайного возмущения, г. с. перейдем от экономической модели к эконометрической. Случайные возмущения включаются в поведенческие уравнения системы и не включаются в уравнения тождества.

В спецификации поведенческим уравнением является первое, таким образом, спецификация *эконометрической модели,* составленная с использованием четырех принципов, следующая:

$$C_t = a + b\,Y_{t-1} + \varepsilon_t, \quad 0<b<1, \quad a>0,$$

$$Y_t = C_t + I_t$$

где ε_t — случайное возмущение, учитывающее влияние не включенных н данное уравнение факторов.

Г. Составим приведенную форму спецификации. Решим систему и выразим эндогенные переменные модели через предопределенные и явном виде. В первом уравнении системы эндогенная переменная уже явно выражена через предопределенную (лаговое значение эндогенной переменной Y_{t-1}), поэтому оставим его без изменения. Подставим первое уравнение системы во второе и выразим текущую эндогенную переменную Y_t через предопределенные переменные модели (Y_{t-1}, I_t):

$$Y_t = a + bY_{t-1} + I_t + \varepsilon_t .$$

Таким образом, приведенная форма модели принимает вид:

$$C_t = a + bY_{t-1} + \varepsilon_t , \; Y_t = a + bY_{t-1} + I_t + \varepsilon_t .$$

Д. *Матричный вид структурной и приведенной форм спецификации.*
Сформируем векторы эндогенных и предопределенных переменных модели.

Вектор-столбец эндогенных переменных модели:

$$Y_t = (C_t, Y_t)^T;$$

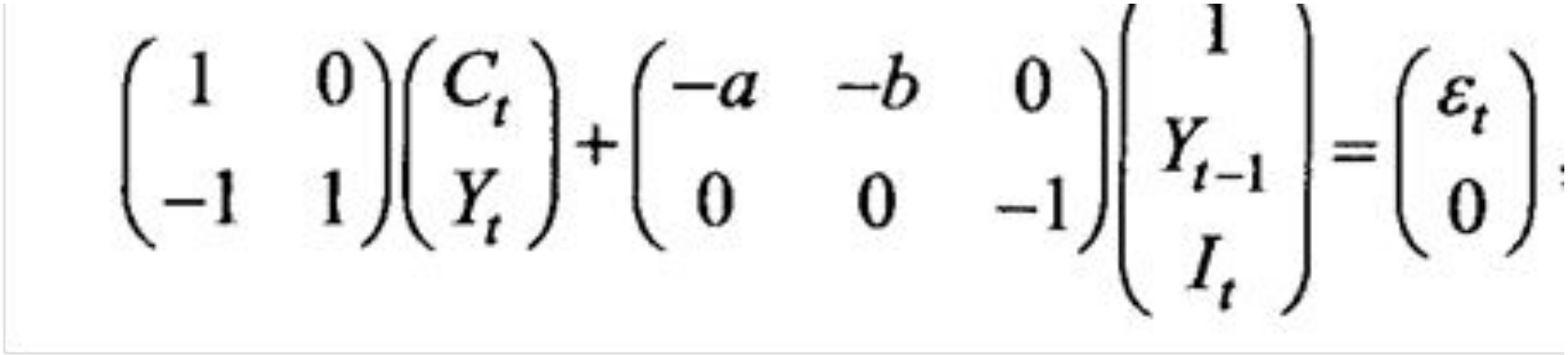
расширенный вектор-столбец предопределенных переменных модели:

$$X_t = (1, Y_{t-1}, I_t)^T .$$

Запишем уравнения структурной формы в следующем виде:

$$C_t - a\ bY_{t-1} = \varepsilon_t,\ Y_t - C_t - I_t = 0\ .$$

тогда матричный вид структурной формы следующий



Решим матричное уравнение относительно вектора эндогенных переменных

$$Y_t = -A^{-1} B X_t + A^{-1} V_t = M X_t + U_t,$$

где $U_t = A^{-1} V_t$.

Обратим матрицу *A,* выразим элементы матрицы *M* через структурные коэффициенты, а вектор возмущений приведенной формы через вектор возмущений структурной формы:

$$A^{-1}=\begin{pmatrix}1 & 0\\ 1 & 1\end{pmatrix},\qquad M=-A^{-1}\cdot B=\begin{pmatrix}-1 & 0\\ -1 & -1\end{pmatrix}\begin{pmatrix}-a & -b & 0\\ 0 & 0 & -1\end{pmatrix}=\begin{pmatrix}a & b & 0\\ a & b & 1\end{pmatrix},$$

$$U_t=\begin{pmatrix}1 & 0\\ 1 & 1\end{pmatrix}\cdot\begin{pmatrix}\varepsilon_t\\ 0\end{pmatrix}=\begin{pmatrix}\varepsilon_t\\ \varepsilon_t\end{pmatrix}.$$

Приведенная форма модели принимает вид

$$\begin{pmatrix}C_t\\ Y_t\end{pmatrix}=\begin{pmatrix}a & b & 0\\ a & b & 1\end{pmatrix}\begin{pmatrix}1\\ Y_{t-1}\\ I_t\end{pmatrix}+\begin{pmatrix}\varepsilon_t\\ \varepsilon_t\end{pmatrix}.$$

Парная линейная регрессия

Сущность регрессионного анализа

Функция регрессии Y на X.

$$M(Y \mid x) = f(x),$$

Где X - независимая (объясняющая) переменная (регрессор),

Y — зависимая (объясняемая) переменная

Множественная регрессия

$$M(Y \mid x1, x2, \ldots, xm) = f(x1, x2, \ldots, xm),$$

Регрессионные модели (уравнения)

$$Y = M(Y \mid x) + \varepsilon,$$

$$Y = M(Y \mid x1, x2, \ldots, xm) + \varepsilon,$$

Причины обязательного присутствия в регрессионных моделях случайного фактора (отклонения)

1. *Невключение в модель всех объясняющих переменных.*
2. *Неправильный выбор функциональной формы модели.*
3. *Агрегирование переменных.*
4. *Ошибки измерений.*
5. *Ограниченность статистических данных.*
6. *Непредсказуемость человеческого*

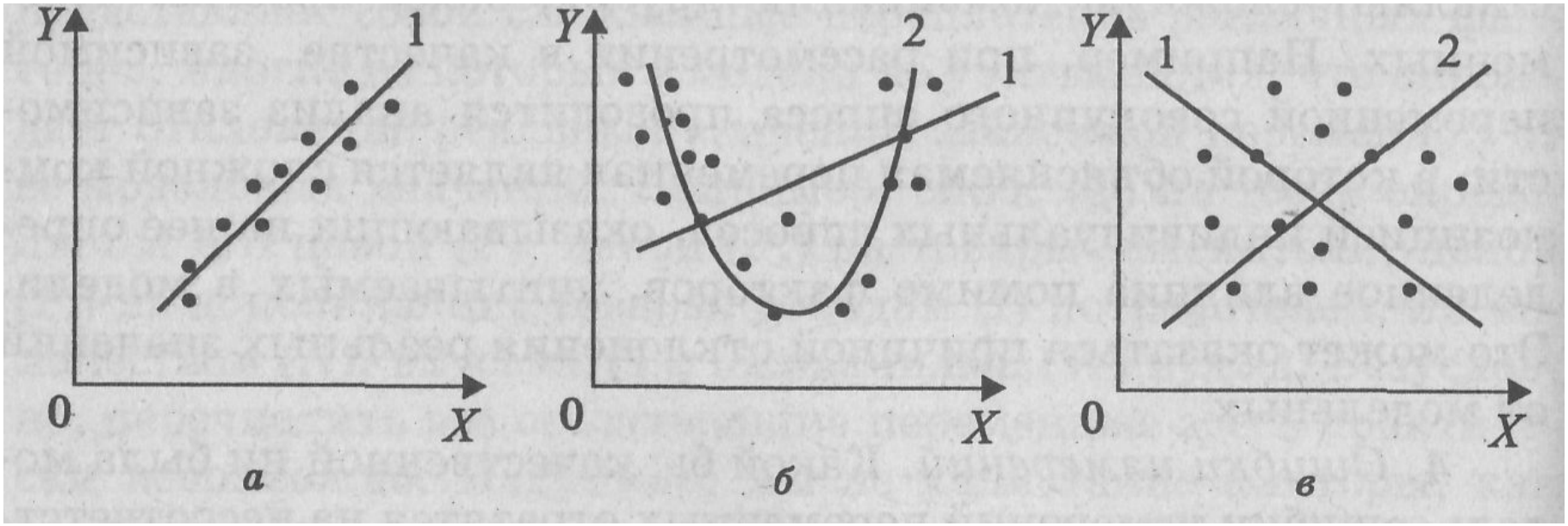
Этапы построения уравнения регрессии

1) выбор формулы уравнения регрессии;

2) определение параметров выбранного уравнения;

3) анализ качества уравнения и проверка адекватности уравнения эмпирическим данным, совершенствование уравнения.

Корреляционное поле (диаграмма рассеивания)



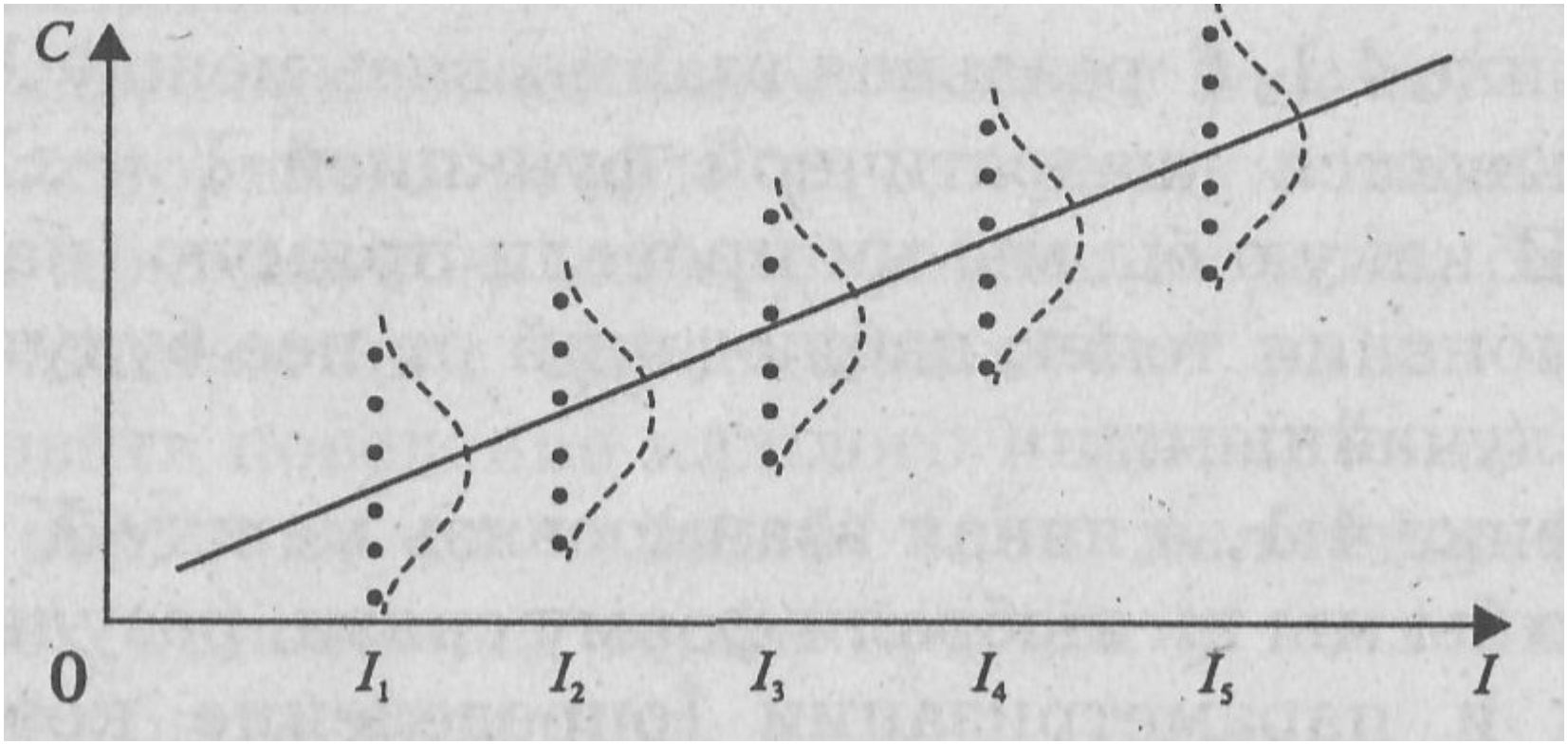
Парная линейная регрессия

Модель Кейнса

$$I = C = C_0 + bI,$$

где C_0 —величина автономного потребления,

b $(0 < b < 1)$ — предельная склонность к потреблению


C
0
I_1
I_2
I_3
I_4
I_5
I

линейная регрессия (теоретическое линейное уравнение регрессии)

$$M(Y \mid x = x_i) = \beta_0 + \beta_1 x_i,$$

Или со случайным параметром ε

$$y_i = M(Y \mid X = x_i) + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

где

β_0 и β_1 — теоретические параметры регрессии;

ε_i — случайное отклонение.

Задачи линейного регрессионного анализа

1. По имеющимся статистическим данным (x_i, y_i), $i = 1, 2, \ldots, n$, для переменных X и Y;

а) получить наилучшие оценки неизвестных параметров β_0 и β_1 ;

б) проверить статистические гипотезы о параметрах модели;

в) проверить, достаточно ли хорошо модель согласуется со статистическими данными (адекватность модели данным наблюдений).

Эмпирическое уравнение регрессии

$$\hat{y}_i = b_0 + b_1 x_1,$$

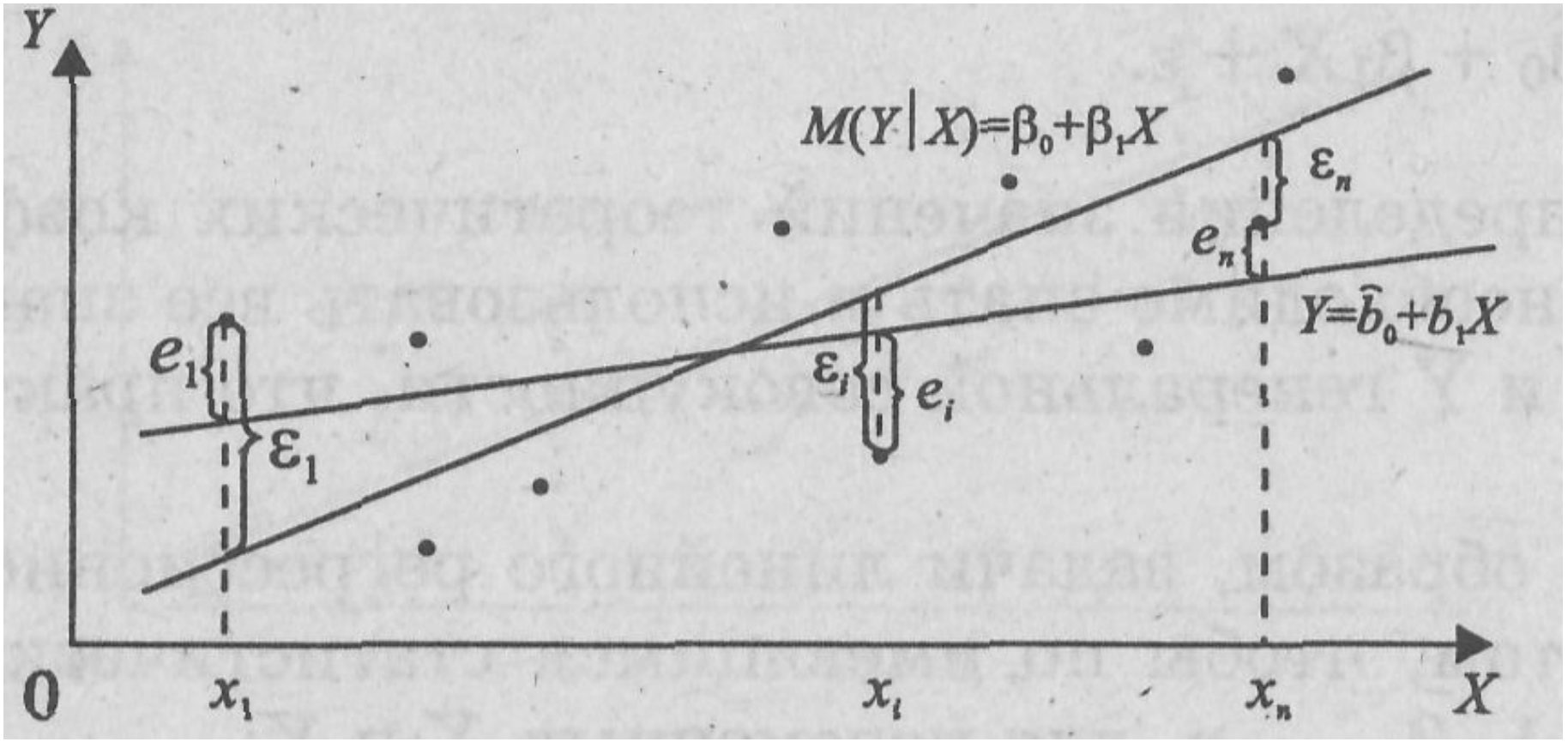
где $\hat{y}_i$ — оценка условного математического ожидания $M(Y \mid X = x_i)$;

b_0 и b_1 — оценки неизвестных параметров β_0 и β_1, называемые *эмпирическими коэффициентами регрессии.*

Тогда

$$y_i = b_0 + b_1 x_i + e_i.$$

где *отклонение* e_i — оценка теоретического случайного отклонения ε_i.


Y
M(Y|X)=β0+β1X
εn
en
Y=b̂0+b1X
e1
ε1
εi
ei
0
x1
xi
xn
X

Оценка тесноты связи

Мерой линейной зависимости двух случайных величин является ковариация этих величин, определяемая выражением

$$\mathbf{Cov(x,y) = S_{xy} = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{n}}$$

$$\sigma_x^2 = S_x^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

$$r = \frac{N\sum_{i=1}^{N} x_i \cdot y_i - \sum_{i=1}^{N} x_i \cdot \sum_{i=1}^{N} y_i}{\sqrt{N \cdot \sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2} \cdot \cdot \sqrt{N \cdot \sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2}}$$

Нахождение коэффициентов b_0 и b_1 эмпирического уравнения регрессии

1) $$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i);$$

2) $$\sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |y_i - \hat{y}_i| = \sum_{i=1}^{n} |y_i - b_0 - b_1 x_i|;$$

3) $$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

Возможные методы нахождения коэффициентов b_0 и b_1 эмпирического уравнения регрессии

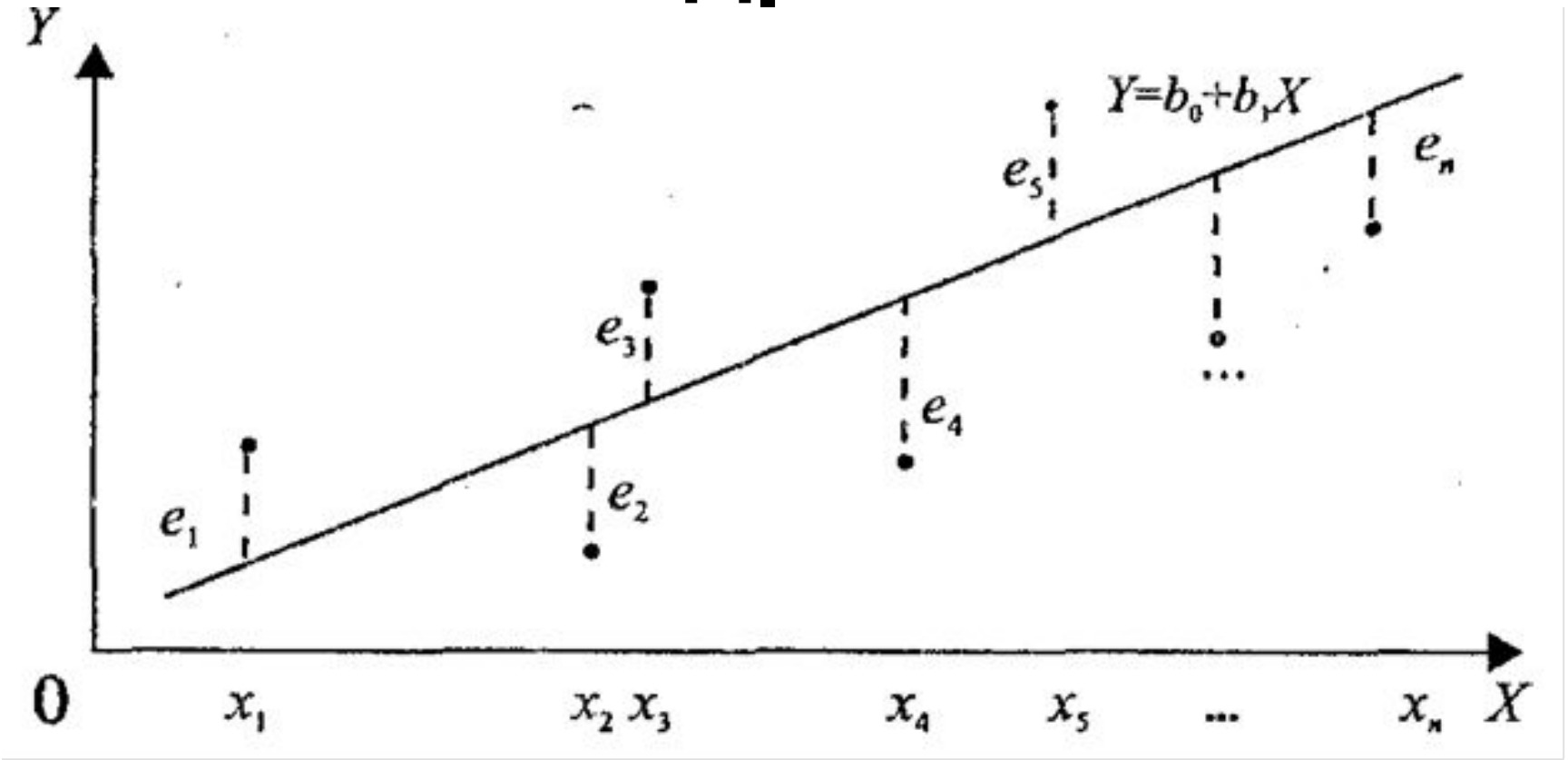
метод наименьших модулей *(МНМ).*

метод наименьших квадратов *(МНК)*

метод моментов (*ММ*)

метод максимального правдоподобия (*ММП*)

Метод наименьших квадратов



$$Q(b_0, b_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

$$\begin{cases} \dfrac{\partial Q}{\partial b_0} = -2\sum (y_i - b_0 - b_1 x_i) = 0; \\ \dfrac{\partial Q}{\partial b_1} = -2\sum (y_i - b_0 - b_1 x_i) x_i = 0. \end{cases} \Rightarrow$$

$$\begin{cases} nb_0 + b_1 \sum x_i = \sum y_i; \\ b_0 \sum x_i + b_1 \sum x_i^2 = \sum x_i y_i. \end{cases}$$

$$\begin{cases} b_0 + b_1\bar{x} = \bar{y}; \\ b_0\bar{x} + b_1\overline{x^2} = \overline{xy}. \end{cases} \Rightarrow \begin{cases} b_1 = \dfrac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - \bar{x}^2}; \\ b_0 = \bar{y} - b_1\bar{x}. \end{cases}$$

$$\bar{x} = \frac{1}{n}\sum x_i,\ \overline{x^2} = \frac{1}{n}\sum x_i^2,\quad \bar{y} = \frac{1}{n}\sum y_i.$$

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_x^2}.$$

Если, кроме уравнения регрессии Y на X ($Y = b_0 + b_1X$), для тех же эмпирических данных найдено уравнение регрессии X на Y ($X = c_0 + b_yY$), то произведение коэффициентов b_x и b_y равно r^2_{xy}:

$$b_x \cdot b_y = r_{xy} \frac{S_y}{S_x} \cdot r_{xy} \cdot \frac{S_x}{S_y} = r_{xy}^2 .$$

$$\begin{cases} b_y = \dfrac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{y^2} - \overline{y}^2}; \\ c_0 = \overline{x} - b_y \overline{y}. \end{cases}$$

Выводы

1. Оценки МНК являются функциями от выборки, что позволяет их легко рассчитывать.
2. Оценки МНК являются точечными оценками теоретических коэффициентов регрессии.
3. Эмпирическая прямая регрессии обязательно проходит через точку $(\bar{x}; \bar{y})$.
4. Эмпирическое уравнение регрессии построено таким образом, что сумма отклонений $\sum e_i$, а также среднее значение от клонения e равны нулю.

5.П Случайные отклонения e_i не коррелированы с наблюдаемыми значениями y_i зависимой переменной *Y*.

6. Случайные отклонения e_i не коррелированы с наблюдаемыми значениями x_i независимой переменной *X*.

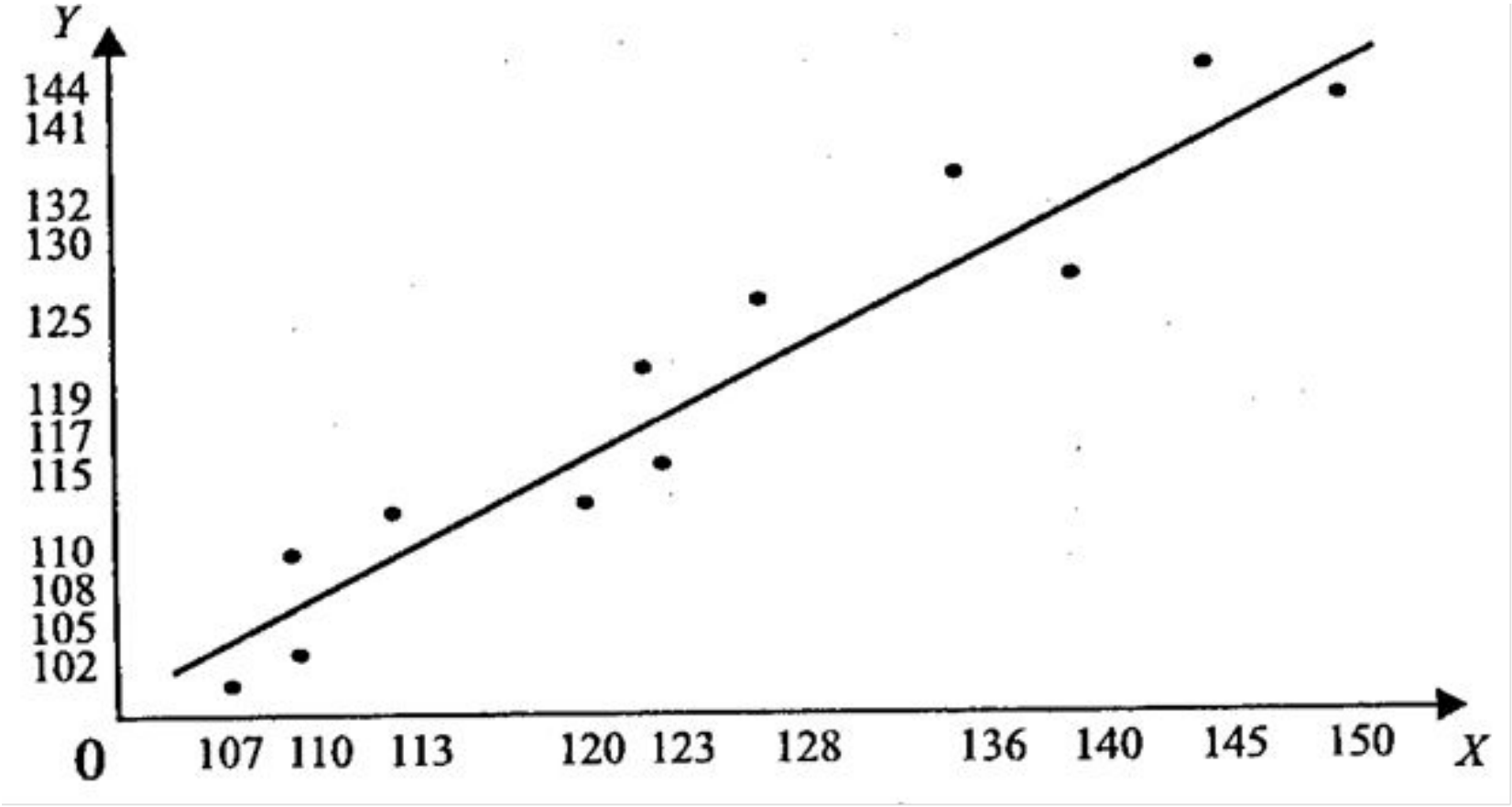
Пример

Для анализа зависимости объема потребления *Y*(у.е.) домохозяйства от располагаемого дохода *X*(у.е.) отобрана вы борка объема *n* = 12 (помесячно в течение года), результаты которой приведены в таблице. Необходимо определить вид зависимости; по МНК оценить параметры уравнения регрессии У на *X;* оценить силу линейной зависимости между *X* и *Y;* спрогнозировать потребление при доходе *X* = 160

Исходные данные

i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	107	109	110	113	120	122	123	128	136	140	145	**150**
y_i	**102**	**105**	**108**	**110**	**115**	**117**	**119**	**125**	**132**	**130**	**141**	**144**

Поле корреляции



i	x_i	y_i	x_i^2	$x_i y_i$	y_i^2	$\hat{y}_i$	e_i	e_i^2
1	107	102	11 449	10 914	10 404	103,63	–1,63	2,66
2	109	105	11 881	11 445	11 025	105,49	–0,49	0,24
3	110	108	12 100	11 880	11 664	106,43	1,57	2,46
4	113	110	12 769	12 430	12 100	109,23	0,77	0,59
5	120	115	14 400	13 800	13 225	115,77	–0,77	0,59
6	122	117	14 884	14 274	13 689	117,63	–0,63	0,40
7	123	119	15 129	14 637	14 161	118,57	0,43	0,18
8	128	125	16 384	16 000	15 625	123,24	1,76	3,10
9	136	132	18 496	17 952	17 424	130,71	1,29	1,66
10	140	130	19 600	18 200	16 900	134,45	–4,45	19,8
11	145	141	21 025	20 445	19 881	139,11	1,89	3,57
12	150	144	22 500	21 600	20 736	143,78	0,22	0,05
Сумма	1503	1448	190 617	183 577	176 834	–	≈ 0 **)	35,3
Среднее*)	125,25	120,67	15884,75	15298,08	14736,17	–	–	–

$$\begin{cases} b_1 = \dfrac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} = \dfrac{15298{,}08 - 125{,}25 \cdot 120{,}67}{15884{,}75 - (125{,}25)^2} = \dfrac{184{,}1625}{197{,}1875} = 0{,}9339; \\ b_0 = \overline{y} - b_1 \overline{x} = 120{,}67 - 0{,}9339 \cdot 125{,}25 = 3{,}699. \end{cases}$$

$$r_{xy} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sqrt{\overline{x^2} - \overline{x}^{\,2}} \cdot \sqrt{\overline{y^2} - \overline{y}^{\,2}}} = \frac{184{,}1625}{14{,}04 \cdot 13{,}23} = 0{,}9914.$$

Выводы

Прогнозируемое потребление при располагаемом доходе x = **160** по данной модели составит $y(160) \approx 153{,}12$.

- коэффициент b_1 может трактоваться как предельная склонность к потреблению (MPC ≈ 0,9339). Фактически он показывает, на какую величину изменится объем потребления, если располагаемый доход возрастает на одну единицу.

- На графике коэффициент b_1 определяет тангенс угла наклона прямой регрессии относительно положительного направления оси абсцисс (объясняющей переменной). Поэтому часто он называется *угловым коэффициентом.*
- *Свободный член* y_0 уравнения регрессии определяет прогнозируемое значение *Y* при величине располагаемого дохода *X,* равной нулю (т.е. автономное потребление).

- Очень важно, насколько далеко данные наблюдений за объясняющей переменной отстоят от оси ординат (зависимой переменной), так как даже при удачном подборе уравнения регрессии для интервала наблюдений нет гарантии, что оно останется таковым и вдали от выборки. В нашем случае значение b_0 = 3,699 говорит о том, что при нулевом располагаемом доходе расходы на потребление составят в среднем 3,699 у.е.

Этот факт можно объяснить для отдельного домохозяйства (оно может тратить накопленные или одолженные средства), но для совокупности домохозяйств он теряет смысл. В любом случае значение коэффициента b_0 определяет точку пересечения прямой регрессии с осью ординат и характеризует сдвиг линии регрессии вдоль оси У.

- Следует помнить, что эмпирические коэффициенты регрессии b_0 и b_1 являются лишь оценками теоретических коэффи циентов β_0 и β_1, а само уравнение отражает лишь общую тенденцию в поведении рассматриваемых переменных. Индивидуальные значения переменных в силу различных причин могут отклоняться от модельных значений. В нашем примере эти отклонения выражены через значения e_i которые являются оценками отклонений ε_i для генеральной совокупности.

Однако при определенных условиях уравнение регрессии служит незаменимым и очень качественным инструментом анализа и прогнозирования. Обсуждение этих условий будет проведено в последующих главах.

- После интерпретации результатов закономерен вопрос о качестве оценок и самого уравнения в целом.

ПРОВЕРКА КАЧЕСТВА УРАВНЕНИЯ РЕГРЕССИИ

Классическая линейная регрессионная модель

Рассмотрим модель парной линейной регрессии

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

$$\tilde{Y} = b_0 + b_1 X$$

Предпосылки метода наименьших квадратов

1 . *Математическое ожидание случайного отклонения ε_i равно нулю: $M(\varepsilon_i) = 0$ для всех наблюдений.*

Данное условие означает, что случайное отклонение в среднем не оказывает влияния на зависимую переменную. В каждом конкретном наблюдении случайный член может быть либо положительным, либо отрицательным, но он не должен иметь систе матического смещения. Отметим, что выполнимость $M(\varepsilon_i) = 0$ влечет выполнимость

2. *Дисперсия случайных отклонений, постоянна:* $D(\varepsilon_i) = D(\varepsilon_j) = \sigma^2$ *для любых наблюдений i и j.*

Данное условие подразумевает, что несмотря на то, что при каждом конкретном наблюдении случайное отклонение может быть либо большим, либо меньшим, не должно быть некой априорной причины, вызывающей большую ошибку (отклонение).

Выполнимость данной предпосылки называется *гомоскедастичностью (постоянством дисперсии отклонений).* Невыполнимость данной предпосылки называется *гетероскедастичностью (непостоянством*

3. *Случайные отклонения* ε_i *и* ε_j *являются независимыми друг от друга для* $i \neq j$.

Выполнимость данной предпосылки предполагает, что отсутствует систематическая связь между любыми случайными отклонениями. Другими словами, величина и определенный знак любого случайного отклонения не должны быть причинами величины и знака любого другого отклонения.

Выполнимость данной предпосылки влечет следующее соотношение:

$$\sigma_{\varepsilon_i \varepsilon_j} = cov(\varepsilon_i, \varepsilon_j) = \begin{cases} 0, & i \neq j \\ \sigma^2, & i = j \end{cases} \qquad (5.6)$$

Поэтому, если данное условие выполняется, то говорят об отсутствии *автокорреляции.* С учетом выполнимости предпосылки 1 соотношение (5.6) может быть переписано в виде: $M(\varepsilon_i, \varepsilon_j) = 0$ $(i \neq j)$

4. *Случайное отклонение должно быть независимо от объясняющих переменных.*

Обычно это условие выполняется автоматически, если объясняющие переменные не являются случайными в данной модели.

5 . *Модель является линейной относительно параметров.*

Теорема Гаусса-Маркова

Если предпосылки 1 — 5 выполнены, то оценки, полученные по МНК, обладают следующими свойствами:

1. Оценки являются несмещенными, т.е. $M(b_0) = \beta_0$, $M(b_1) = \beta_1$. Это вытекает из того, что $M(\varepsilon_i) = 0$, и говорит об отсутствии систематической ошибки в определении положения линии регрессии.

2. Оценки состоятельны, так как дисперсия оценок параметров при возрастании числа *n* наблюдений стремится к нулю. Другими словами, при увеличении объема выборки надежность оценок увеличивается (b_0 наверняка близко к β_0, b_1 — близко к β_1).

3. Оценки эффективны, т.е. они имеют наименьшую дисперсию по сравнению с любыми другими оценками данных па раметров, линейными относительно величин y_i.

Анализ точности определения оценок коэффициентов регрессии

- Модель парной линейной регрессии

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

- Пусть на основе выборки из *n* наблюдений оценивается регрессия

$$\tilde{Y} = b_0 + b_1 X$$

- где

$$b_1 = \frac{S_{xy}}{S_x^2}$$

То есть коэффициент b_1 также является случайным, так как значение выборочной ковариации S_{xy} зависит от того, какие значения принимают *X* и *Y.* Если *X* можно рассматривать как экзогенный фактор, значения которого известны, то значения *Y* зависят от случайной составляющей ε_i.

Теоретически коэффициент b_1 можно разложить на неслучайную и случайную составляющие.

$$S_{xy} = \text{cov}\,(X, \beta_0 + \beta_1 X + \varepsilon) = \text{cov}(X, \beta_0) + \text{cov}\,(X, \beta_1 X) + \text{cov}\,(X, \varepsilon) \Rightarrow$$

$$S_{xy} = \beta_1 + \text{cov}\,(X, \varepsilon).$$

Здесь использовались правила вычисления ковариации:

$$\text{cov}(X, \beta_0) = 0, \text{ так как } \beta_0 = \text{const},$$

$$\text{cov}(X, \beta_1 X) = \beta_1 \text{cov}(X, X) = S_x^2.$$

Следовательно,

$$b_1 = \frac{S_{xy}}{S_x^2} = \beta_1 + \frac{S_{x\varepsilon}}{S_x^2}$$

Здесь β_1 — постоянная величина (истинное значение коэффициента регрессии);

$\frac{S_{x\varepsilon}}{S_x^2}$ — случайная компонента.

Аналогичный результат можно получить и для коэффициента b_0.

Вывод

На практике такое разложение осуществить невозможно, поскольку неизвестны истинные значения β_0 и β_1, а также значения отклонений для всей генеральной совокупности.

Найдем формулы связи дисперсий коэффициентов $D(b_0)$ и $D(b_1)$ с дисперсией σ^2 случайных отклонений ε_i.

Для этого представим формулы определения коэффициентов b_0 и b_1 в виде линейных функций относительно значений Y:

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} - \frac{\bar{y} \sum (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \Rightarrow$$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}, \text{ так как } \sum (x_i - \bar{x}) = 0.$$

имеем:

$$b_1 = \sum c_i y_i$$

$$b_0 = \sum d_i y_i$$

где

$$c_i = \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

$$d_i = \frac{1}{n} - c_i \bar{x}$$

$$D(b_1) = D(\sum c_i y_i) = \sigma^2 \sum c_i^2 = \frac{\sigma^2}{\sum (x_i - \bar{x})^2},$$

$$D(b_0) = D(\sum d_i y_i) = \sigma^2 \sum d_i^2 = \sigma^2 \sum \left(\frac{1}{n} - c_i \bar{x} \right)^2 =$$

$$= \sigma^2 \sum \left(\frac{1}{n^2} - \frac{2 c_i \bar{x}}{n} + c_i^2 \bar{x}^2 \right)^2 =$$

$$= \sigma^2 \left(\frac{1}{n} - 0 + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right) =$$

$$= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right) = \frac{\sigma^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}.$$

выводы

- Дисперсий b_0 и b_1 прямо пропорциональны дисперсии случайного отклонения σ^2. Следовательно, чем больше фактор случайности, тем менее точными будут оценки.
- Чем больше число n наблюдений, тем меньше дисперсии оценок. Это вполне логично, так как чем большим числом данных мы располагаем, тем вероятнее получение более точных оценок.

- Чем больше дисперсия (разброс значений) объясняющей переменной, тем меньше дисперсия оценок ко эффициентов. Другими словами, чем шире область изменений объясняющей переменной, тем точнее будут оценки (тем меньше доля случайности в их определении).

В силу того что случайные отклонения ε_i по выборке определены быть не могут, при анализе надежности оценок коэффициентов регрессии они заменяются отклонениями

$$e_i = y_i - b_0 - b_1 x_i$$

значений y_i переменной Y от оцененной линии регрессии.

Дисперсия случайных отклонений $D(\varepsilon_i) = \sigma^2$ заменяется ее несмещенной оценкой

$$S^2 = \frac{1}{n-2}\sum\left(y_i - b_0 - b_1 x_i\right)^2 = \frac{\sum e_i^2}{n-2}$$

Тогда

$$D(b_1) \approx S_{b_1}^2 = \frac{S^2}{\sum\left(x_i - \overline{x}\right)^2}$$

$$D(b_0) \approx S_{b_0}^2 = \frac{S^2 \sum x_i^2}{n\sum\left(x_i - \overline{x}\right)^2} = \overline{x^2} S_{b_1}^2$$

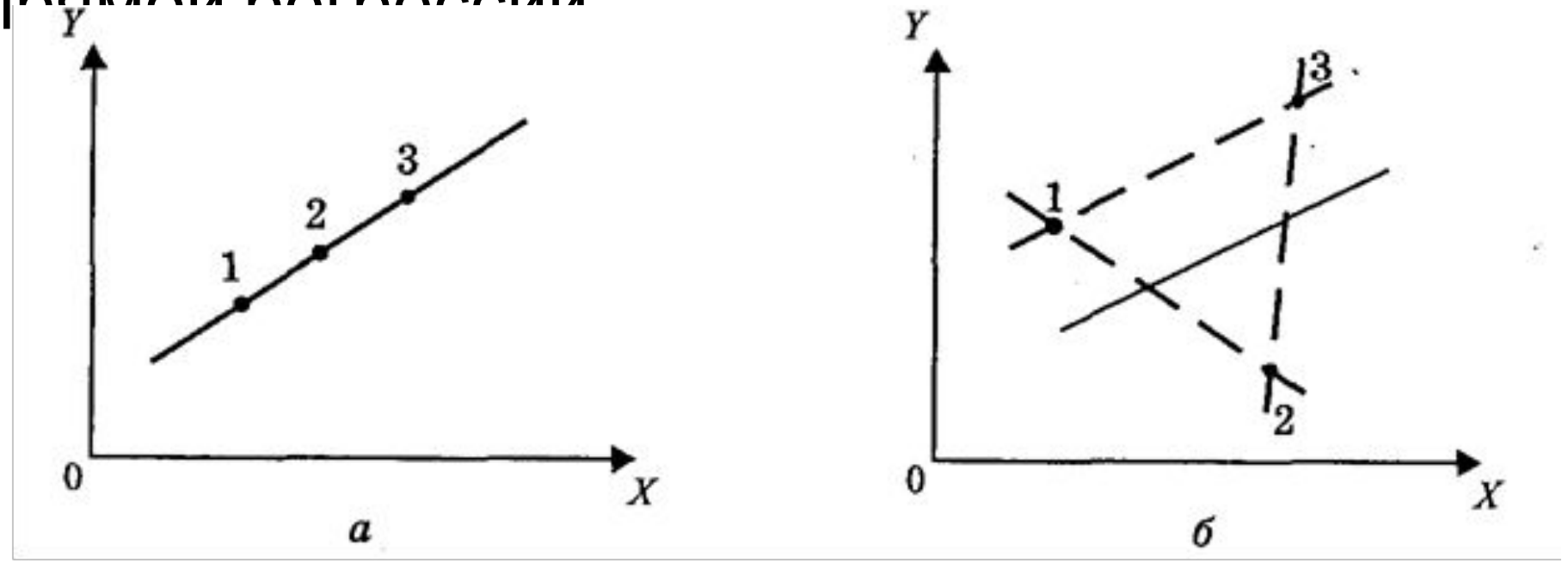
Где $S^2 = \frac{\sum e_i^2}{n-2}$ - необъясненная дисперсия (мера разброса зависимой переменной вокруг линии регрессии).

Корень квадратный из необъясненной дисперсии называется *стандартной ошибкой оценки (стандартной ошибкой регрессии).*

$S_{b_1}^2$ $S_{b_0}^2$ — стандартные отклонения случайных величин b_0 и b_1, называемые *стандартными ошибками коэффициентов регрессии.*

Графическая интерпретация

Коэффициент b_1 определяет наклон прямой регрессии. Чем больше разброс значений Y вокруг линии регрессии, тем больше (в среднем) ошибка определения наклона прямой регрессии.

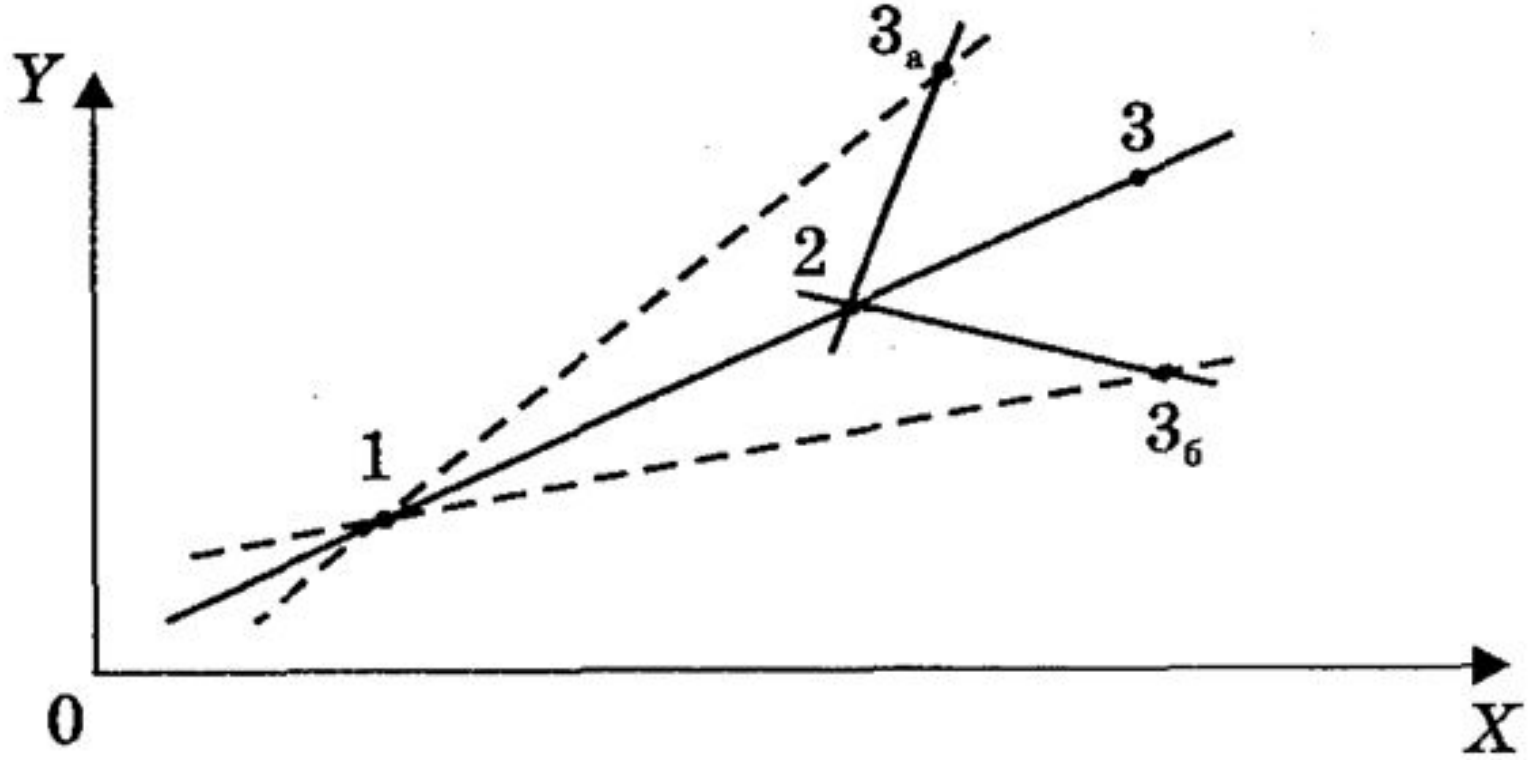


Например, на рис. а все наблюдаемые точки лежат на одной прямой. Тогда через любой набор точек проводится одна и та же прямая.

На рис. *б* точки не лежат на одной прямой, но для трех точек прямая регрессии будет такой же (хотя отклонения от линии регрессии существенны), как и на рис. *а.* Однако при исключении из рассмотрения любой из указанных трех точек прямые регрессии будут существенно отличаться друг от друга ((1, 2), (1, 3), (2, 3)). Следовательно, значительно различаются их углы наклона, а значит, стандартная ошибка коэффициента

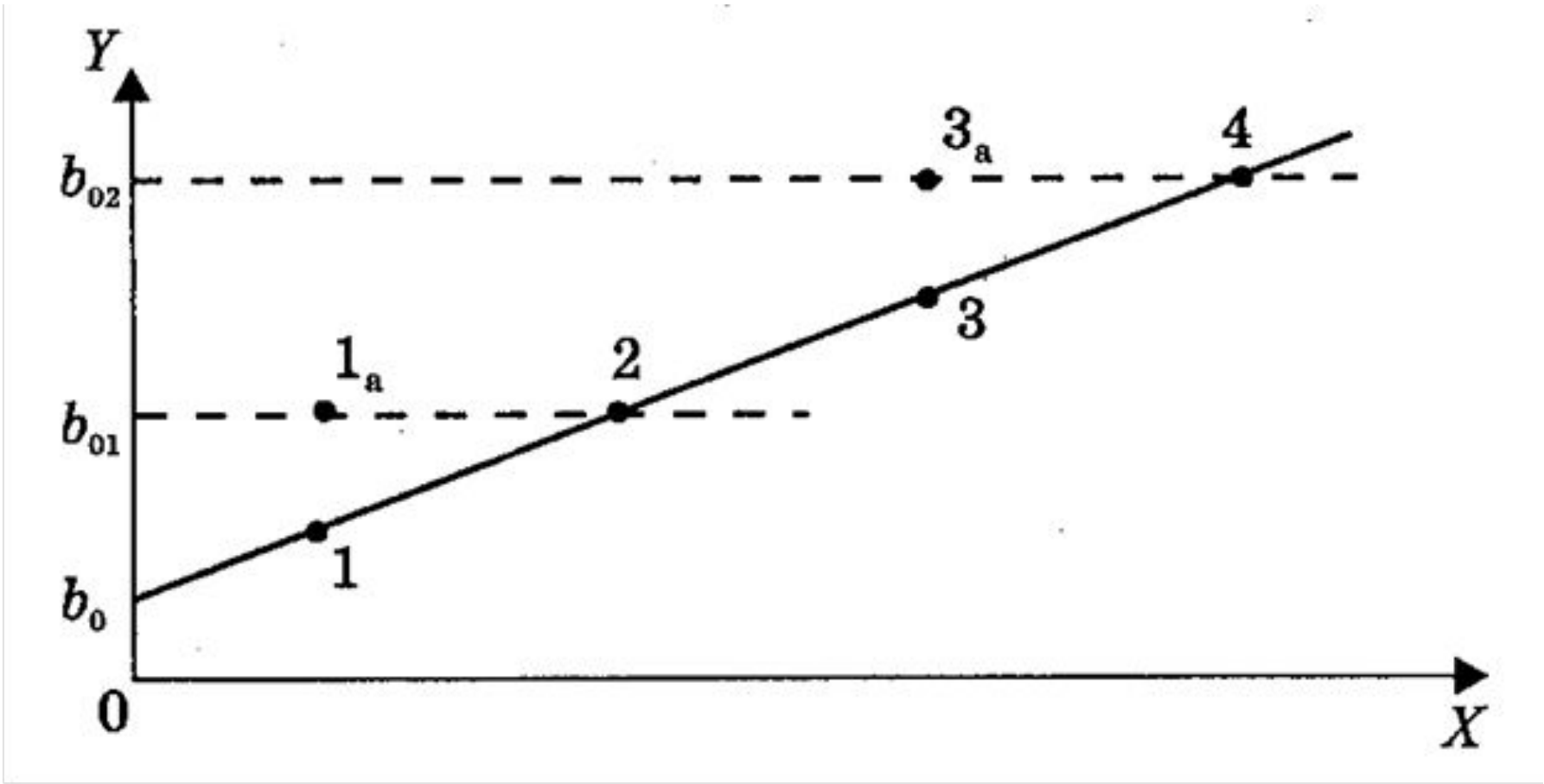
В выражении, определяющим значение *стандартной ошибки коэффициента регрессии* b_1, стоит сумма квадратов отклонений x_i от среднего значения. Эта сумма велика (а следовательно, вся дробь мала, и дисперсия оценки меньше), если регрессия определяется на широком диапазоне зна чений переменной *X*.

Например, на рис. через пары точек (1, 3) и (2, 3) проведена одна и та же прямая. Но диапазон (1, 3) шире диапазона (2, 3). Если вместо точки 3 рассмотреть либо точку $3_{а}$, либо $3_{б}$ (т.е. при случайном изменении выборки), то наклон прямой для пары (1, 3) изменится значительно меньше, чем для пары (2, 3).


Y
0
X
1
2
3
3а
3б

Дисперсия свободного члена уравнения регрессии пропорциональна дисперсии коэффициента регрессии.

Действительно, чем сильнее меняется наклон прямой, проведенной через данную точку, тем большее разброс значений свободного члена, характеризующего точку пересечения этой прямой с осью $0Y$.


Y
b_{02}
b_{01}
b_0
0
X
3_a
4
3
1_a
2
1

На рис. через пары точек (1, 2) и (3, 4) проходит одна и та же прямая, пересекающая ось O*Y* в точке (0, b_0). Для второй из этих пар значения переменной *X* больше по абсолютной ве личине (при одинаковом диапазоне изменений *X* и *Y*), чем для первой. Если в этих парах точки 1 и 3 изменить на одну и ту же величину (новые точки 1_a, 3_a), то углы наклона новых прямых (1_a, 2) и (3_a, 4) будут одинаковы. Но свободный член b_{01} для первой прямой будет существенно меньше отличаться от b_0, чем свободный член b_{02} для второй прямой.

Проверка гипотез относительно коэффициентов линейного уравнения регрессии

Эмпирическое уравнение регрессии определяется на основе конечного числа статистических данных. Поэтому коэффициенты эмпирического уравнения регрессии являются СВ, изменяющимися от выборки к выборке.

При проведении статистического анализа перед исследователем зачастую возникает необходимость сравнения эмпирических коэффициентов регрессии b_0 и b_1 с некоторыми теоретически ожидаемыми значениями β_0 и β_1 этих коэффициентов.

Данный анализ осуществляется по схеме статистической проверки гипотез.

Статистической называют гипотезу о виде закона распределения или о параметрах известного распределения. В первом случае гипотеза называется *непараметрической,* а во втором — *параметрической.*

- Гипотеза H_0, подлежащая проверке, называется *нулевой (основной).*
- Наряду с нулевой рассматривают гипотезу H_1, которая будет приниматься, если отклоняется H_0. Такая гипотеза называется *альтернативной (конкурирующей).*

Например, если проверяется гипотеза о равенстве параметра θ некоторому значению θ_0, т.е. H_0: $\theta = \theta_0$, то в качестве альтернативной могут рассматриваться следующие гипотезы:

$$H_1^{(1)}: \theta \neq \theta_0;\ H_1^{(2)}: \theta > \theta_0;\ H_1^{(3)}: \theta < \theta_0;\ \ H_1^{(4)}: \theta = \theta_1\ (\theta_1 \neq \theta_0).$$

Гипотезу называют *простой,* если она содержит одно конкретное предположение

$$H_0 : \theta = \theta_0,\ H_1^{(4)} : \theta = \theta_1$$

Гипотезу называют *сложной,* если она состоит из конечного или бесконечного числа простых гипотез

$$H_1^{(1)} : \theta \neq \theta_0;\ H_1^{(2)} : \theta > \theta_0;\ H_1^{(3)} : \theta < \theta_0$$

- При проверке гипотезы выборочные данные могут противоречить гипотезе H_0. Тогда она *отклоняется.*
- Если же статистические данные согласуются с выдвинутой гипотезой, то она *не отклоняется.* В последнем случае часто говорят, что нулевая гипотеза принимается (такая формулировка не совсем точна, однако она широко распространена).
- Статистическая проверка гипотез на основании выборочных данных неизбежно связана с риском принятия ложного решения.

При этом возможны ошибки двух родов:

- *Ошибка первого рода* состоит в том, что будет отвергнута правильная нулевая гипотеза.
- *Ошибка второго рода* состоит в том, что будет принята нулевая гипотеза, в то время как в действительности верна альтернативная гипотеза.

результаты статистических выводов

Результаты проверки гипотезы	Возможные состояния гипотезы	
	верна H_0	верна H_1
Гипотеза $H_„$ отклоняется	Ошибка первого рода	Правильный вывод
Гипотеза H_0 не отклоня ется	Правильный вывод	Ошибка второго рода

Вероятность совершить ошибку первого рода принято обозначать буквой α, и ее называют *уровнем значимости.* Вероятность совершить ошибку второго рода обозначают β. Тогда вероятность не совершить ошибку второго рода (1 — β) называется *мощностью критерия.*

Проверку статистической гипотезы осуществляют на основании данных выборки. Для этого используют специально подобранную случайную величину (статистику, критерий), точное или приближенное значение которой известно.

Наиболее известные случайные величины (статистики, критерии)

- *U* (или *Z)* — стандартизированное нор мальное распределение;
- *T* — если она распределена по закону Стьюдента;
- — если она распределена по закону ;
- *F* — если она имеет распределение Фишера.

В целях общности будем обозначать такую случайную величину через *К.*

Основной принцип проверки статистических гипотез можно сформулировать так: если наблюдаемое значение критерия *К* (вычисленное по выборке) принадлежит критической области, то нулевую гипотезу отклоняют. Если же наблюдаемое значение критерия *К* принадлежит области принятия гипотезы, то нулевую гипотезу не отклоняют (принимают).

- Точки, разделяющие критическую область и область принятия гипотезы, называют

Вероятность того, что случайная величина *K* попадет в произвольный интервал ($k_{1-\alpha/2}$ $k_{\alpha/2}$), можно найти по формуле

$$P\left(k_{1-\alpha/2} < K < k_{\alpha/2}\right) = \int_{1-\alpha/2}^{\alpha/2} f\left(k\middle|H_0\right)dk$$

Зададим эту вероятность равной 1 — α и вычислим критические точки (квантили) *K*-распределения

$$P\left(K < k_{1-\alpha/2}\right) = \int_{-\infty}^{k_{1-\alpha/2}} f\left(k\middle|H_0\right)dk = \frac{\alpha}{2}$$

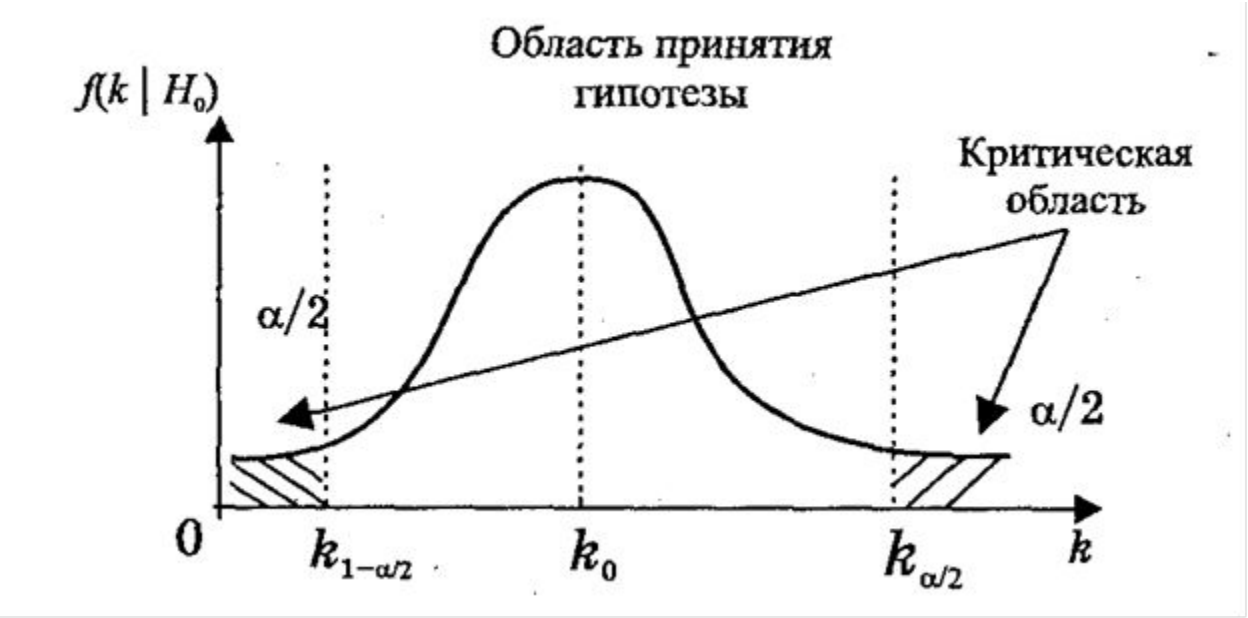
$$P\left(K \geq k_{\alpha/2}\right) = \int_{k_{\alpha/2}}^{\infty} f\left(k\middle|H_0\right)dk = \frac{\alpha}{2}$$

Следовательно

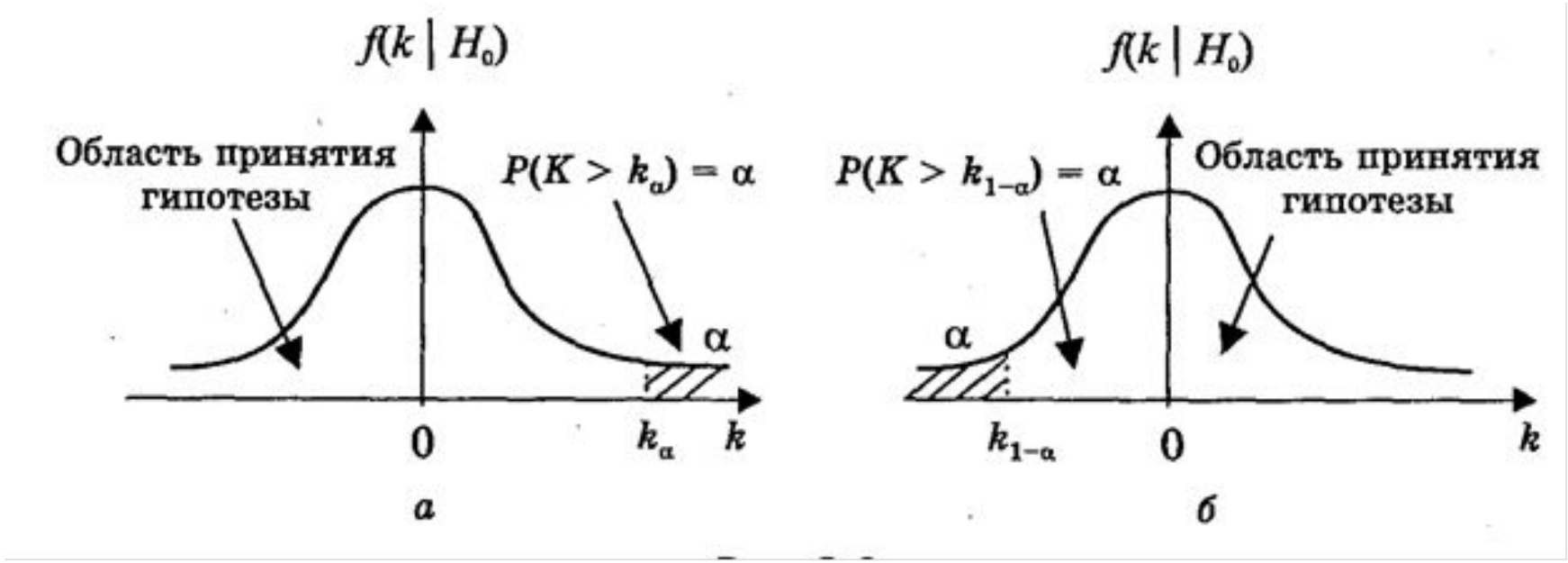
$$P\left(k_{1-\alpha/2} < K < k_{\alpha/2}\right) = 1 - \alpha$$

$$P\left(\left(K \le k_{1-\alpha/2}\right) < \cup \left(K \ge k_{\alpha/2}\right)\right) = \alpha$$

- *двусторонней критической областью*



односторонняя критическая область — правосторонняя или *левосторонняя*



Общая схема проверки гипотез

1. Формулировка проверяемой (нулевой — H_0) и альтернативной (H_1) гипотез.
2. Выбор соответствующего уровня значимости α.
3. Определение объема выборки *п.*
4. Выбор критерия *К* для проверки H_0.
5. Определение критической области и области принятия гипотезы.
6. Вычисление наблюдаемого значения критерия $K_{набл}$.
7. Принятие статистического решения.

Проверка гипотез и доверительные интервалы

Для проверки гипотезы

$$H_0 : b_1 = \beta_1,\ H_1 : b_1 \neq \beta_1,$$

используется статистика

$$t = \frac{b_1 - \beta_1}{S_{b_1}}$$

которая при справедливости H_0 имеет распределение Стьюдента с числом степеней свободы v = n — 2, где n — объем выборки.

Следовательно, H_0: $b_1 = \beta_1$ отклоняется на основании данного критерия, если

$$\left|T_{\text{набл}}\right| = \left|\frac{b_1 - \beta_1}{S_{b_1}}\right| \geq t_{\frac{\alpha}{2}, n-2}$$

где α — требуемый уровень значимости.

При невыполнении этого соотношения считается, что нет оснований для отклонения H_0.

- Наиболее важной на начальном этапе статистического анализа построенной модели все же является задача установле ния наличия линейной зависимости между *Y* и *X*. Эта проблема может быть решена по той же схеме:

$$H_0 : b_1 = 0, \; H_1 : b_1 \neq 0.$$

Гипотеза в такой постановке обычно называется *гипотезой о статистической значимости коэффициента регрессии.* При этом, если H_0 принимается, то есть основания считать, что величина *Y* не зависит от *X*.

В этом случае говорят, что коэффициент b_1 *статистически незначим* (он слишком близок к нулю).

При отклонении H_0 коэффициент b_1 считается *статистически значимым,* что указывает на наличие определенной линейной зависимости между *Y* и *X.*

В данном случае рассматривается двусторонняя критическая область, так как важным является именно отличие от нуля коэффициента регрессии, и он может быть как положительным, так и отрицательным.

Поскольку полагается, что $\beta_1 = 0$, то формально значимость оцененного коэффициента регрессии b_1 проверяется с помощью анализа отношения его величины к его стандартной ошибке .

При выполнении исходных предпосылок модели эта дробь имеет распределение Стьюдента с числом степеней свободы v = *n* — 2, где *n* — число наблюдений. Данное отношение называется *t-статистикой:*

$$t = \frac{b_1}{S_{b_1}} = \frac{b_1}{\sqrt{S_{b_1}^2}}$$

Для *t*-статистики проверяется нулевая гипотеза о равенстве ее нулю.

Очевидно, $t = 0$ равнозначно $b_1 = 0$, поскольку *t* пропорциональна b_1.

Фактически это свидетельствует об отсутствии линейной связи между *X* и *Y*.

Гетероскедастичность

- *Предпосылки МНК (условия Гаусса—Маркова)*

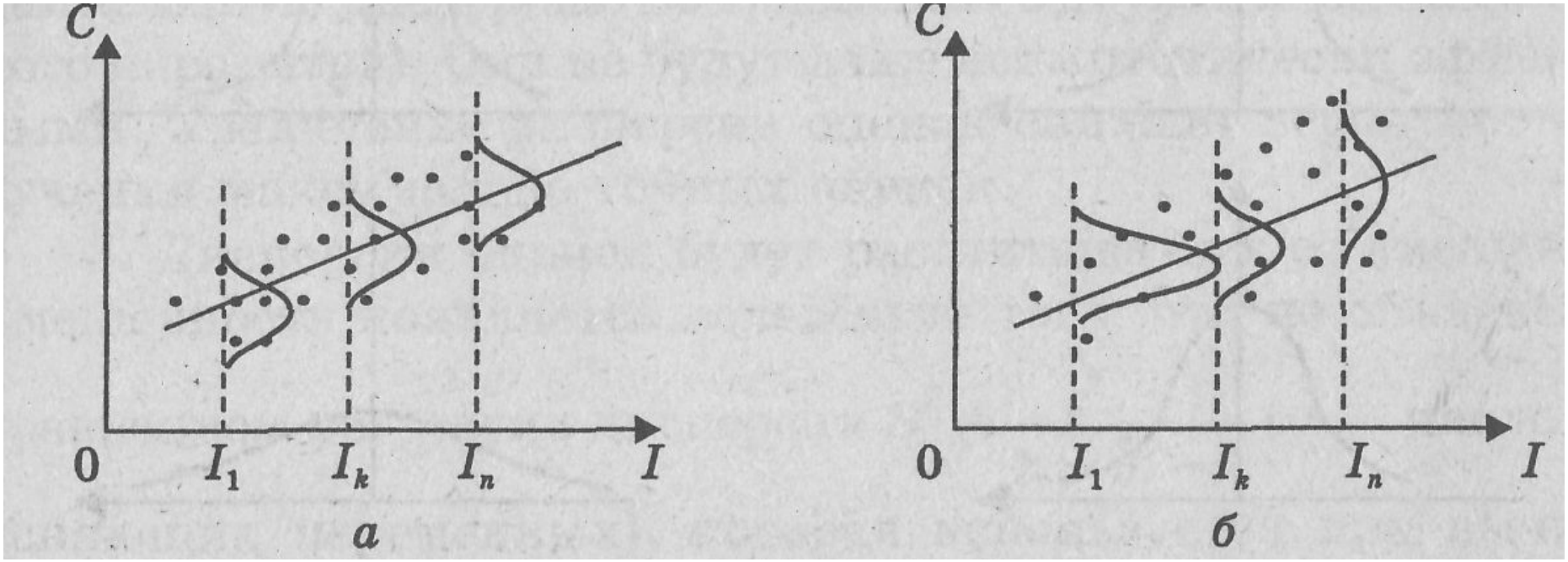
2°. Дисперсия случайных отклонений εi постоянна:

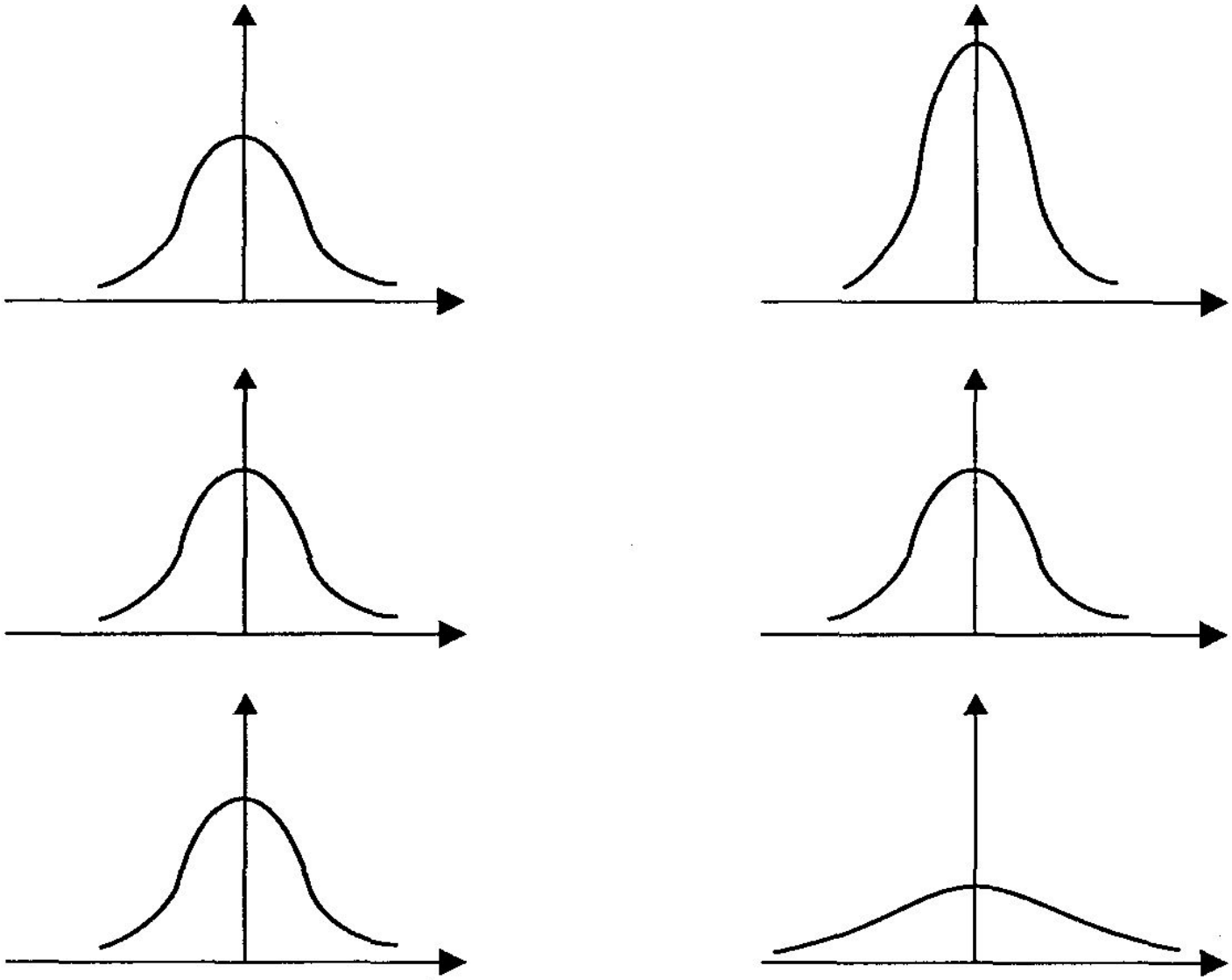
$$D(\varepsilon i) = D(\varepsilon j) = a$$

для любых наблюдений i и j.

Выполнимость данной предпосылки называется гомоскедастичностью (постоянством дисперсии отклонений). Невыполнимость данной предпосылки называется гетероскедастичностью (непостоянством дисперсий отклонений).

- Данное условие подразумевает, что, несмотря на то, что при каждом конкретном наблюдении случайное отклонение может быть большим либо маленьким, положительным либо отрицательным, не должно быть априорной причины, вызывающей большую ошибку (отклонение) при одних наблюдениях и меньшую — при других.


C
0
I_1
I_k
I_n
I
а
C
0
I_1
I_k
I_n
I
б



а *б*

Последствия гетероскедастичности

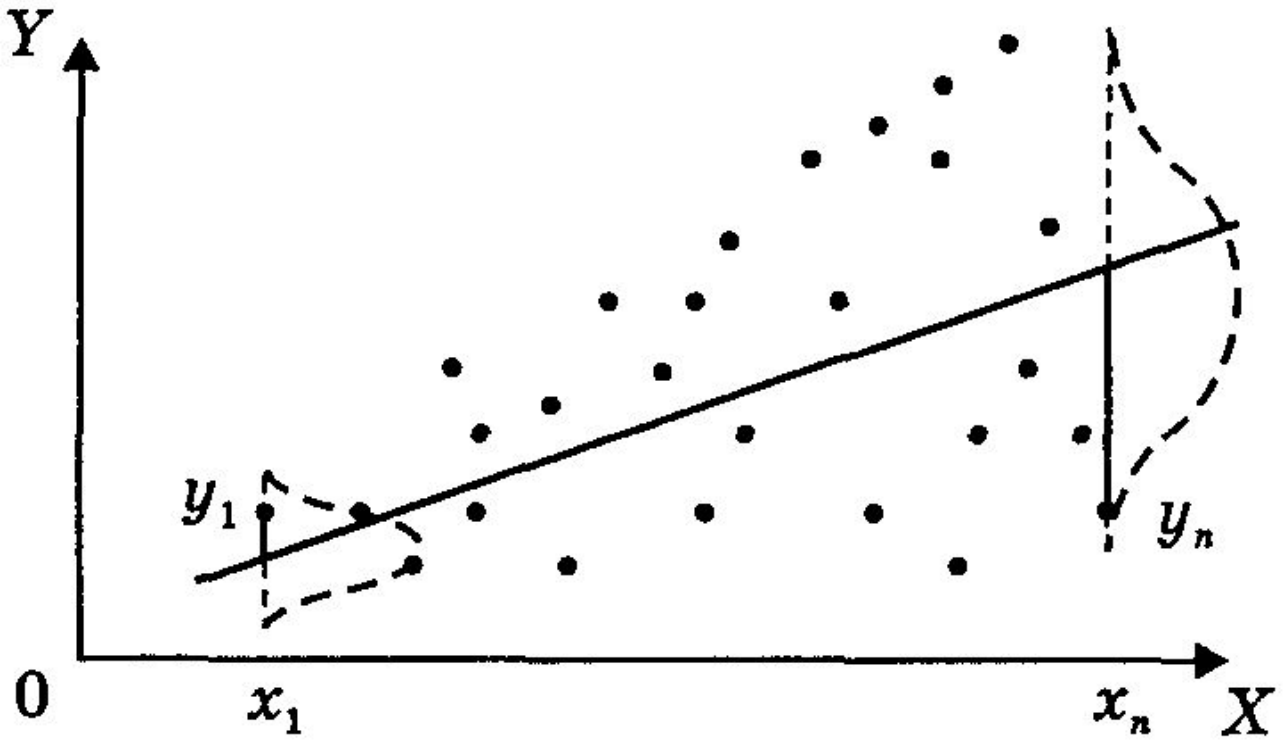
- При гетероскедастичности последствия применения МНК будут следующими.

1. Оценки коэффициентов по-прежнему останутся несмещенными и линейными.
2. Оценки не будут эффективными (т.е. они не будут иметь наименьшую дисперсию по сравнению с другими оценками данного параметра). Они не будут даже асимптотически эффективными. Увеличение дисперсии оценок снижает вероятность получения максимально точных оценок.

3. Дисперсии оценок будут рассчитываться со смещением. Смещенность появляется вследствие того, что не объясненная уравнением регрессии дисперсия, которая используется при вычислении оценок дисперсий всех коэффициентов, не является более несмещенной.

4. Вследствие вышесказанного все выводы, получаемые на основе соответствующих *t*- и *F*-статистик, а также интервальные оценки будут ненадежными. Следовательно, статистические выводы, получаемые при стандартных проверках качества оценок, могут быть ошибочными и приводить к неверным заключениям по построенной модели. Вполне вероятно, что стандартные ошибки коэффициентов будут занижены, а следовательно, *t*-статистики будут завышены. Это может привести к признанию статистически значимыми коэффициентов, таковыми на самом деле не являющихся.

Причина неэффективности оценок МНК при гетероскедастичности



Обнаружение гетероскедастичности

Не существует какого-либо однозначного метода определения гетероскедастичности. Рассмотрим наиболее популярные и наглядные:

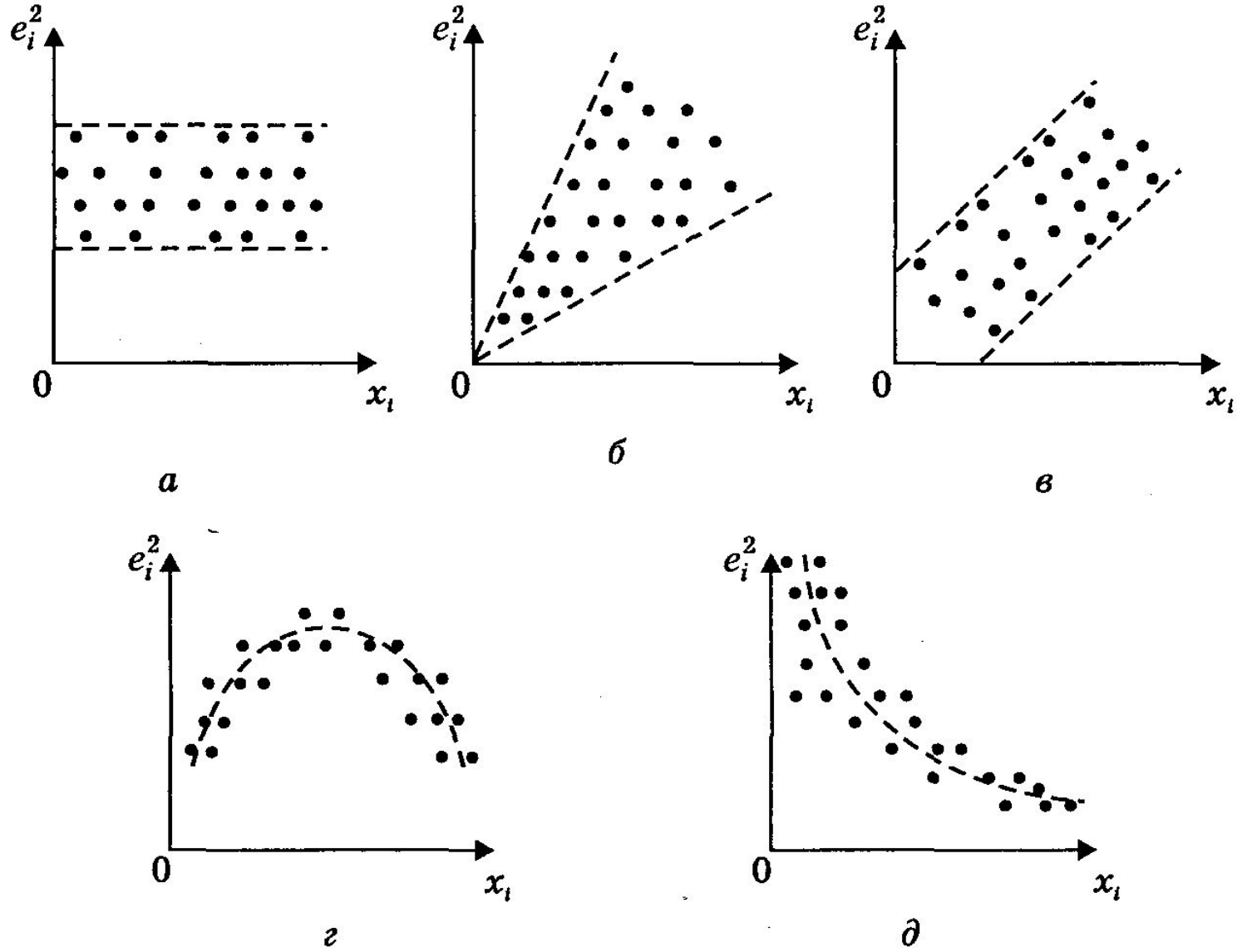
- графический анализ отклонений,
- тест ранговой корреляции Спирмена,
- тест Парка,
- тест Глейзера,
- тест Голдфелда—Квандта.

Графическое представление отклонений

- по оси абсцисс откладываются значения (x_i) объясняющей переменной X (либо линейной комбинации объясняющих переменных

$$Y = b_0 + b_1 X_1 + \ldots + b_m X_m,$$

- а по оси ординат либо отклонения e_i, либо их квадраты e_i^2, $i = 1, 2, \ldots, n$.


e_i^2
0
x_i
а
б
в
г
д

- На рис. *а* все отклонения e_i^2 находятся внутри полуполосы постоянной ширины, параллельной оси абсцисс. Это говорит о независимости дисперсий e_i^2 от значений переменной *X* и их постоянстве, т.е. в этом случае выполняются условия гомоскедастичности.

- На рис. *б—д* наблюдаются некоторые систематические изменения в соотношениях между значениями переменной *X* и квадратами отклонений . e_i^2
- Ситуации, представленные на рис. *б — д,* отражают большую вероятность наличия гетероскедастичности для рассматриваемых статистических данных.

Тест ранговой корреляции Спирмена

Предположение

Дисперсия отклонения будет либо увеличиваться, либо уменьшаться с увеличением значений X.

Поэтому для регрессии, построенной по МНК, абсолютные величины отклонений e_i и значения x_i случайной величины X будут коррелированы.

- Ранжируем, то есть упорядочиваем по величинам значения x_i и e_i .

- Определяем коэффициент ранговой корреляции:

$$r_{x,e} = 1 - 6\frac{\sum d_i^2}{n\left(n^2 - 1\right)}$$

где d_i — разность между рангами x_i и e_i, i = 1, 2, ... , n; n — число наблюдений.

- Если коэффициент корреляции r_{xe} для генеральной совокупности равен нулю, то статистика

$$t = \frac{r_{xe}\sqrt{n-2}}{\sqrt{1-r_{xe}^2}}$$

имеет распределение Стьюдента с числом степеней свободы $v=n-2$.

- Следовательно, если наблюдаемое значение *t*-статистики, превышает $t_{кр} = t_{\alpha/2,n-2}$ (определяемое по таблице критических точек распределения Стьюдента), то необходимо отклонить гипотезу о равенстве нулю коэф фициента корреляции r_{xie}, а следовательно, и об отсутствии гетероскедастичности.
- В противном случае гипотеза об отсутствии гетероскедастичности принимается.

- Если в модели регрессии больше чем одна объясняющая переменная, то проверка гипотезы может осуществляться с помощью *i*-статистики для каждой из них отдельно

Тест Парка

- Критерий Парка дополняет графический метод некоторыми формальными зависимостями.
- Предполагается, что дисперсия

$$\sigma_i^2 = \sigma^2(e_i)$$

является функцией *i*-го значения x_i объясняющей переменной.

- Парк предложил следующую функциональную зависимость:

$$\sigma_i^2 = \sigma^2 x_i^{\beta} e^{vi}$$

Прологарифмировав это выражение, получим:

$$\ln\sigma_i^2 = \ln\sigma^2 + \beta \ln x_i + v_i.$$

Так как дисперсии σ_i^2 обычно неизвестны, то их заменяют оценками квадратов отклонений e_i^2 .

- Критерий Парка включает следующие этапы:

1. Строится уравнение регрессии $y_i = b_0 + b_1 x_i + e_i$.
2. Для каждого наблюдения определяются

$$\ln e_i^2 = \ln(y_i - \hat{y}_i)^2.$$

3. Строится регрессия $\ln e_i^2 = \alpha + \beta \ln x_i + v_i$,

где $\alpha = \ln \sigma^2$.

В случае множественной регрессии такая зависимость строится для каждой объясняющей переменной.

4. Проверяется статистическая значимость коэффициента β уравнения регрессии на основе t-статистики . Если коэффициент β статистически значим, то это означает наличие связи между $\ln e_i^2$ и $\ln x_i$, т.е. гетероскедастичности в статистических данных.

- Отметим, что использование в критерии Парка конкретной функциональной зависимости может привести к необосно ванным выводам (например, коэффициент β статистически незначим, а гетероскедастичность имеет место).

- Возможна еще одна проблема. Для случайного отклонения v_i в свою очередь может иметь место гетероскедастичность. Поэтому критерий Парка дополняется другими тестами.

Тест Глейзера

- Тест Глейзера по своей сути аналогичен тесту Парка и дополняет его анализом других (возможно, более подходящих) зависимостей между дисперсиями отклонений σ_i и значениями переменной x_i. По данному методу оценивается регрессионная зависимость модулей отклонений $|e_i|$ (тесно связанных с σ_i) от x_i.

- Зависимость моделируется следующим уравнением регрессии:

$$|e_i| = \alpha + \beta x_i^k + \nu_i$$

- Изменяя значения *k,* можно построить различные регрессии. Обычно *k* = …, -1, -0,5, 0,5, 1, …
- Статистическая значимость коэффициента β в каждом конкретном случае фактически означает наличие гетероскедастичности. Если для нескольких регрессий коэффициент β оказывается статистически значимым, то при определении характера зависимости обычно ориентируются на

Тест Голдфелда—Квандта

Стандартное отклонение $\sigma_i = \sigma_i(\varepsilon_i)$ пропорционально значению x_i переменной X в этом наблюдении, т.е. , $i = 1, 2, \ldots, n$. Предпо лагается, что ε_i имеет нормальное распределение и отсутствует автокорреляция остатков.

Тест Голдфелда—Квандта состоит в следующем:

1. Все n наблюдений упорядочиваются по

2. Вся упорядоченная выборка после этого разбивается на три подвыборки размерностей *k, (п - 2k), k* соответственно.

3. Оцениваются отдельные регрессии для первой подвыборки *(k* первых наблюдений) и для третьей подвыборки *(k* последних наблюдений). Если предположение о пропорциональности дисперсий отклонений значениям *X* верно, то дисперсия регрессии по первой подвыборке будет существенно меньше дисперсии регрессии по третьей подвыборке.

4. Для сравнения соответствующих дисперсий строится следующая *F*-статистика:

$$F = \frac{\sigma_3 / (k - m - 1)}{\sigma_1 / (k - m - 1)} = \frac{\sigma_3}{\sigma_1}$$

Здесь (*k* - *m* - 1) — число степеней свободы соответствующих выборочных дисперсий *(m* — количество объясняющих переменных в уравнении регрессии).

- При сделанных предположениях относительно случайных отклонений построенная F-статистика имеет распределение Фишера с числами степеней свободы $v_1 = v_2 = k - m - 1$.

5. Если $F_{набл} > F_{табл}$, то гипотеза об отсутствии гетероскедастичности отклоняется при выбранном уровне значимости α.

Методы смягчения проблемы гетероскедастичности

Метод взвешенных наименьших квадратов (ВНК)

- Данный метод применяется при известных для каждого наблюдения значениях дисперсии случайных отклонений . В этом случае можно устранить гетеро скедастичностъ, разделив каждое наблюдаемое значение на со ответствующее ему значение среднего квадратического отклонения. В этом суть метода взвешенных наименьших квадратов.

- Имеем уравнение парной регрессии

$$y_i = \beta_0 + b_1 x_1 + \varepsilon_i$$

Разделим обе части этого уравнения на известное

$$\sigma_i = \sqrt{\sigma_i^2}$$

$$\frac{y_i}{\sigma_i} y_i = \beta_0 \frac{1}{\sigma_i} + b_1 \frac{x_i}{\sigma_i} + \frac{\varepsilon_i}{\sigma_i}$$

Обозначив $\frac{y_i}{\sigma_i} = y_i^*;\ \frac{1}{\sigma_i} = z_i;\ \frac{x_i}{\sigma_i} = x_i^*;\ \frac{\varepsilon_i}{\sigma_i} = v_i$ получим

$$y_i^* = \beta_0 z_i + b_1 x_i^* + v_i$$

- При этом для v_i выполняется условие гомоскедастичности.
- Следовательно, для преобразованной модели выполняются предпосылки МНК. В этом случае оценки, полученные по МНК, будут наилучшими линейными несме щенными оценками.

Этапы ВНК

1. Значения каждой пары наблюдений (x_i, y_i) делят на известную величину среднего квадратического отклонения. Тем самым наблюдениям с наименьшими дисперсиями придаются наибольшие «веса», а с максимальными дисперсиями — наименьшие «веса».

Действительно, наблюдения с меньшими дисперсиями отклонений будут более значи мыми при оценке коэффициентов регрессии, чем наблюдения с большими дисперсиями. Учет этого факта увеличивает вероятность получения более точных оценок.

2. По МНК для преобразованных значений строится уравнение регрессии без свободного члена с гарантированными качествами оценок.

Дисперсии отклонений неизвестны

Для применения ВНК необходимо знать фактические значения дисперсий σ_i^2 отклонений. На практике такие значения известны крайне редко. Следовательно, чтобы применить ВНК, необходимо сделать реалистические предположения о значениях σ_i^2.

- Например, может оказаться целесообразным предположить, что дисперсии σ_i^2 отклонений ε_i; пропорциональны значениям x_i или значениям x_i^2

1. Дисперсии σ_i^2 пропорциональны x_t :

$$\sigma_i^2 = \sigma^2 x_i$$

где σ^2 — коэффициент пропорциональности).

Тогда уравнение регрессии преобразуется делением его левой и правой

$$\frac{y_i}{\sqrt{x_i}} = \beta_0 \frac{1}{\sqrt{x_i}} + b_1 \frac{x_i}{\sqrt{x_i}} + \frac{\varepsilon_i}{\sqrt{x_i}} \Rightarrow \frac{y_i}{\sqrt{x_i}} = \beta_0 \frac{1}{\sqrt{x_i}} + b_1 \sqrt{x_i} + v_i$$

- Для случайных отклонений v_i выполняется условие гомоскедастичности. Следовательно, для принятого уравнения регрессии применим обычный МНК.
- Таким образом, оценив для последнего уравнения по МНК коэффициенты β_0 и β_1 затем возвращаются к исходному уравнению регрессии.

Дисперсии σ_i^2 пропорциональны x^2

- В случае, если зависимость σ_i^2 от x_i целесообразнее выразить не линейной функцией, а квадратичной, то соответствующим преобразованием будет деление уравнения регрессии на x_i:

$$\frac{y_i}{x_i} y_i = \beta_0 \frac{1}{x_i} + b_1 \frac{x_i}{x_i} + \frac{\varepsilon_i}{x_i} \Rightarrow \frac{y_i}{x_i} y_i = \beta_0 \frac{1}{x_i} + b_1 + v_i$$

Для отклонений v_i будет выполняться условие гомоскедастичности.

После определения по МНК оценок коэффициентов β_0 и β_1 для преобразованного уравнения возвращаются к исходному уравнению.

Для применения описанных выше преобразований весьма значимы знания об истинных значениях дисперсий отклонений σ_i^2, либо предположения, какими эти дисперсии могут быть. Во многих случаях дисперсии отклонений зависят не от включенных в уравнение регрессии объясняющих переменных, а от тех, которые не включены в модель, но играют существенную роль в исследуемой зависимости. В этом случае они должны быть включены в модель. В ряде случаев для устранения гетероскедастичности необходимо изменить спецификацию модели.