



МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Истинная модель парной линейной регрессии $Y = a + \mathbf{b} * X + e.$

Для ее оценки используется выборка:

$$(Y_1, X_1)$$

.....

$$(Y_n, X_n)$$

Получается выборочное уравнение регрессии


$$Y = a + b * X$$

Для элемента (Y_i, X_i) выборки, $i = 1, \dots, n$, можно записать:

$$\hat{Y}_i = a + b * X_i$$

$$e_i = Y_i - \hat{Y}_i$$

$$Y_i = a + b * X_i + e_i$$

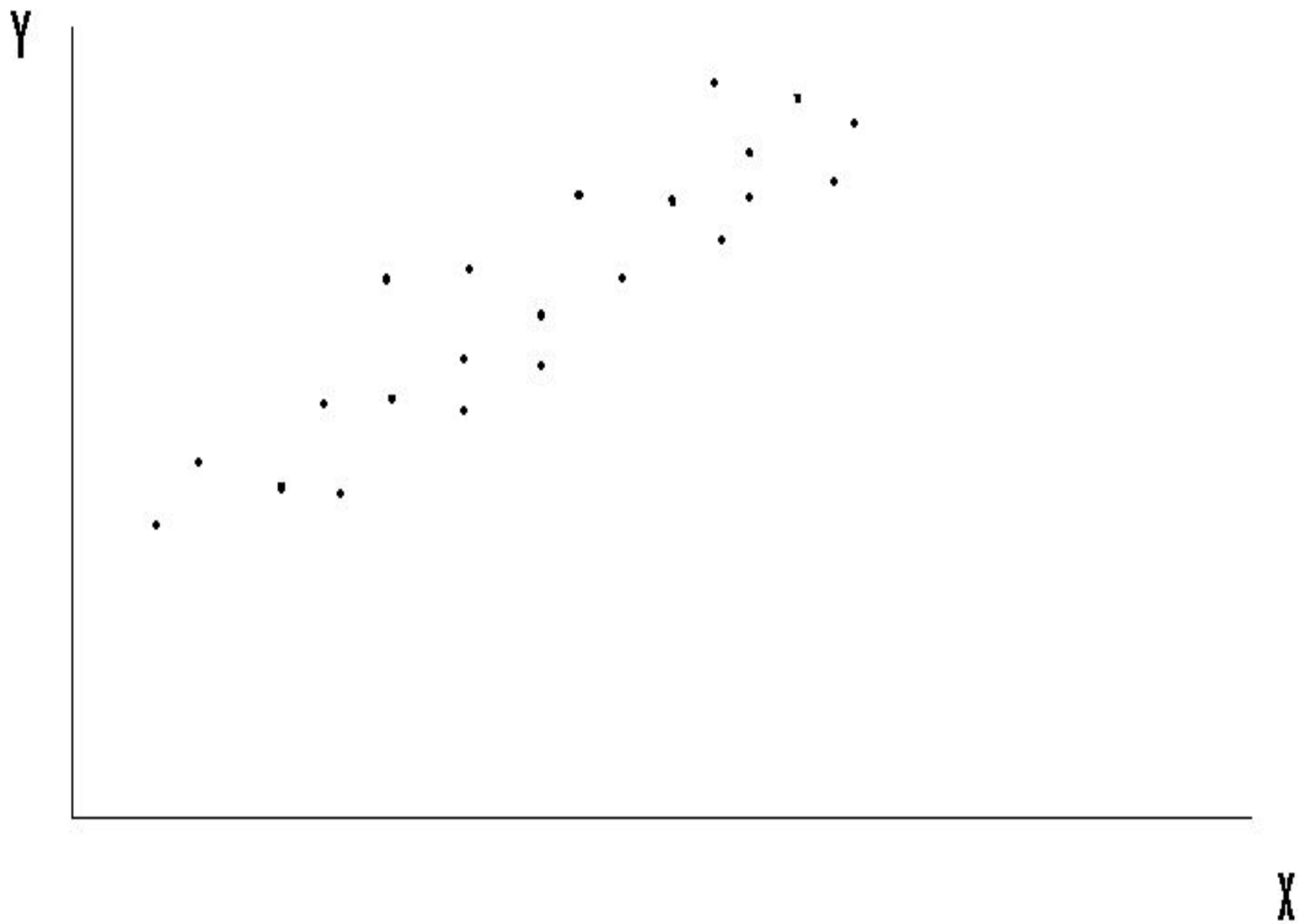


Как оцениваются по выборке
коэффициенты регрессии?

Выборка (Y_i, X_i) , по которой мы должны оценить теоретическую модель

$$Y = a + \mathbf{b} * X + e,$$

графически представляется в виде «облачка» точек:



По этим точкам мы хотим получить такое выборочное уравнение

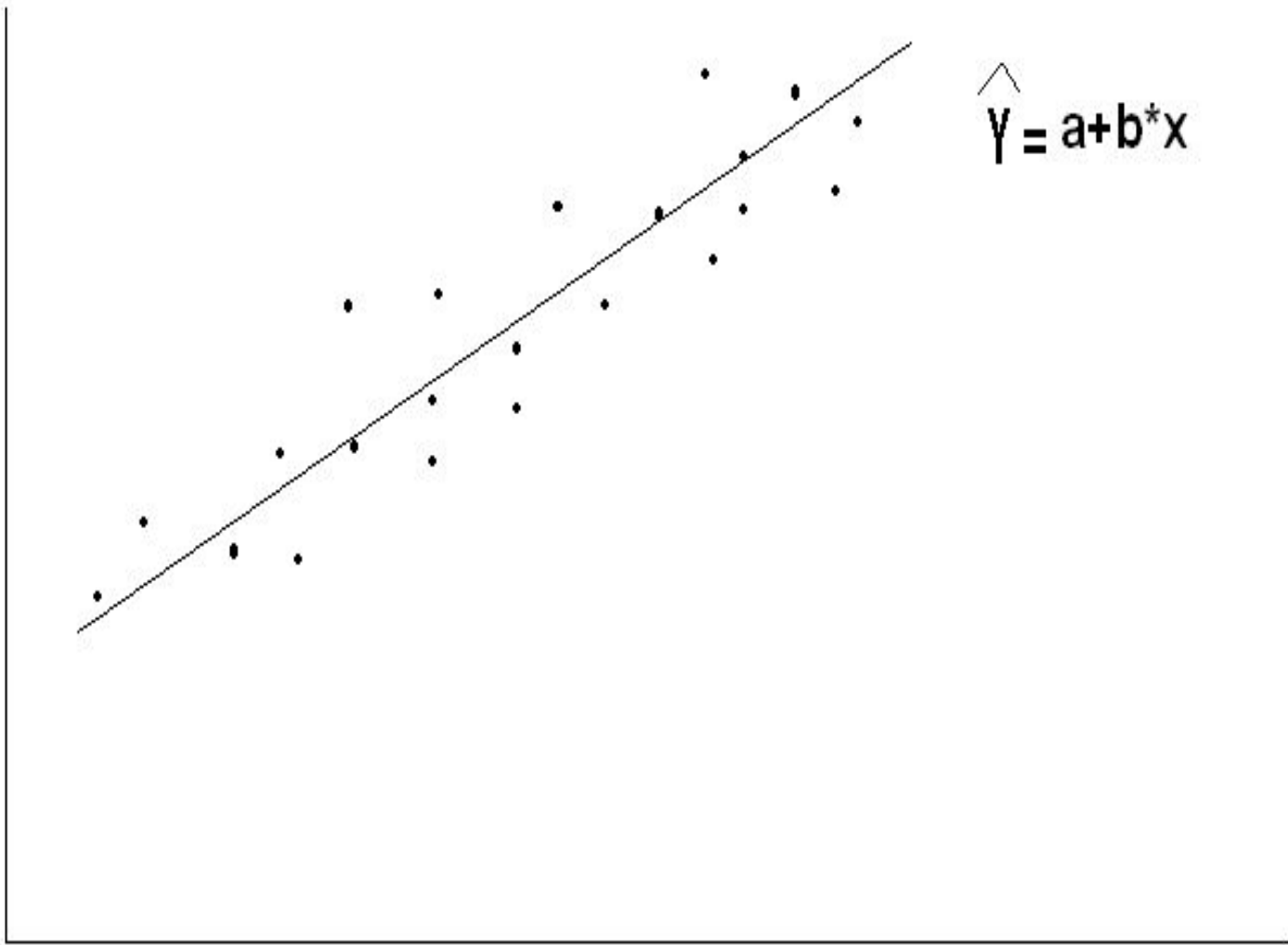
$$Y = a + b * X$$

(т. е. оценки a и b), которое как можно точнее представляло бы истинную линию регрессии

Интуиция подсказывает:

Чем лучше оцененная прямая регрессии представляет выборку, тем точнее она приближает истинную прямую регрессии.

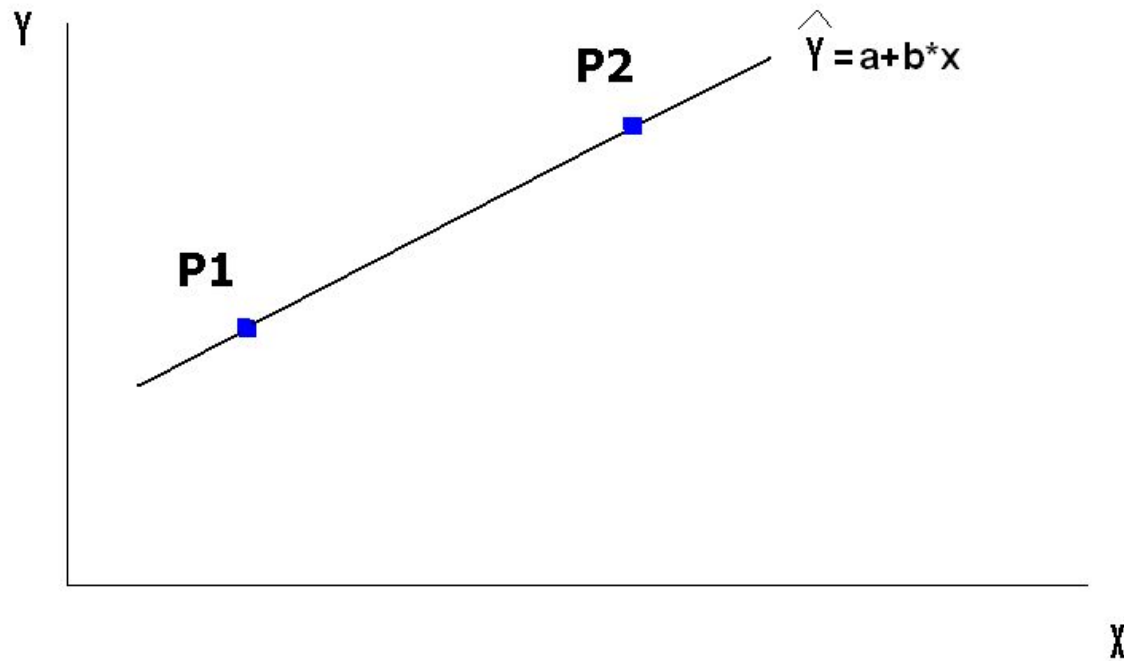
Y



$$\hat{Y} = a + b \cdot X$$

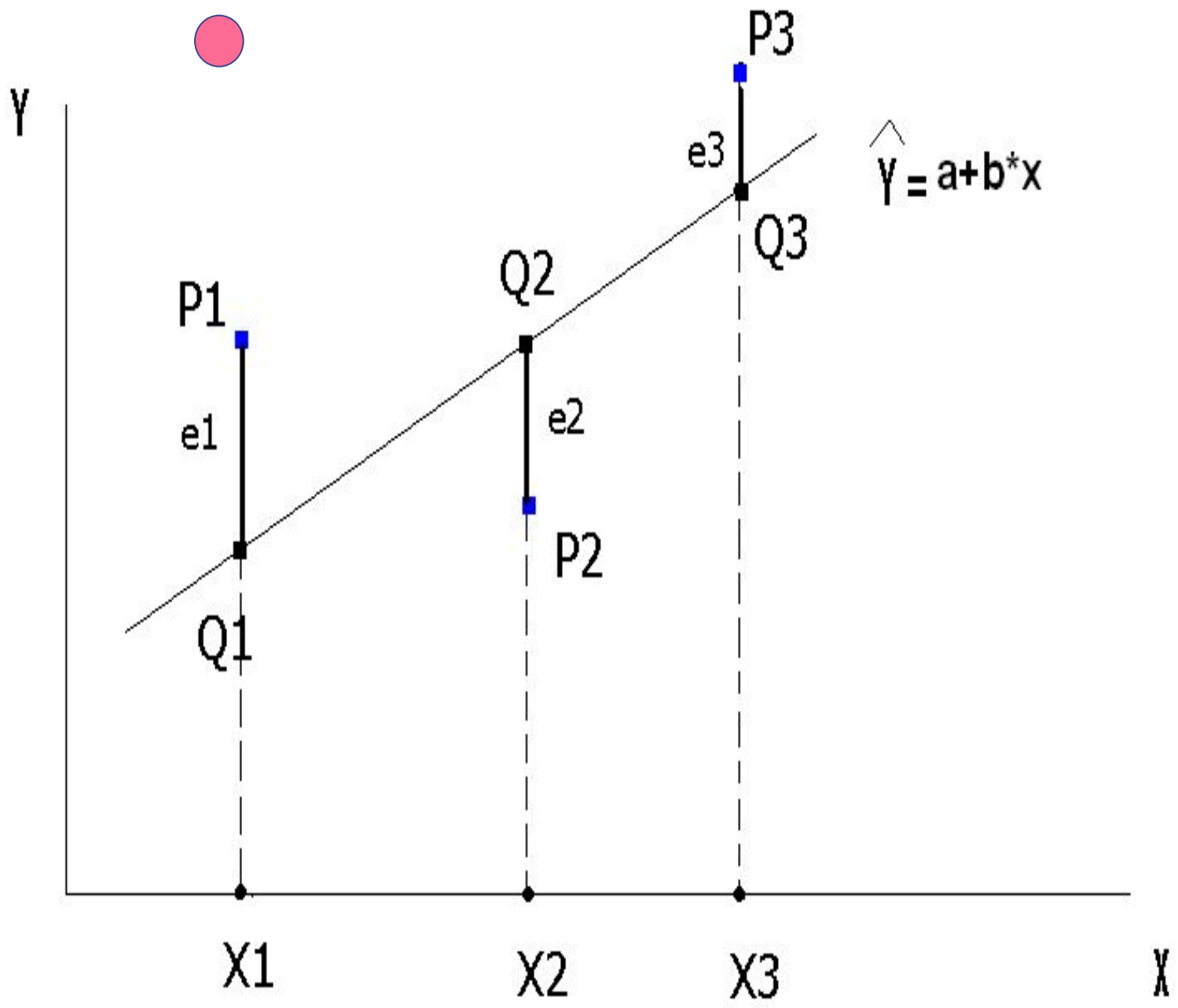
X

Если выборка состоит только из двух точек, то проблем нет:



Если точек больше двух:





Точки выборки

$$P1 = (X_1, Y_1), \quad P2 = (X_2, Y_2), \quad P3 = (X_3, Y_3)$$

моделируются (оцениваются) точками линии регрессии

$$Q1 = (X_1, \hat{Y}_1), \quad Q2 = (X_2, \hat{Y}_2), \quad Q3 = (X_3, \hat{Y}_3).$$

Точность моделирования Y_i для каждого X_i определяется величиной ошибки

$$e_i = Y_i - \hat{Y}_i.$$

Хотелось бы, чтобы выборочное уравнение $\hat{Y} = a + b \cdot X$ с наименьшими ошибками моделировало бы сразу все выборочные значения $Y_i, \quad i = 1, \dots, n.$

Принцип метода наименьших квадратов

Для данной выборки $(X_1, Y_1), \dots, (X_n, Y_n)$ параметры a и b рассчитываются таким образом, чтобы получить минимальное значение суммы квадратов остатков:

$$\min \sum_{i=1}^n e_i^2$$

Или

$$\min \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = S$$

Решаем систему уравнений:

$$\begin{cases} \frac{\partial S}{\partial a} = 0 \\ \frac{\partial S}{\partial b} = 0 \end{cases}$$

После преобразований получаем систему нормальных уравнений для коэффициентов регрессии

$$\begin{cases} n * a - \sum_{i=1}^n Y_i + b * \sum_{i=1}^n X_i = 0 \\ a * \sum_{i=1}^n X_i^2 - \sum_{i=1}^n X_i * Y_i + a * \sum_{i=1}^n X_i = 0 \end{cases}$$

Решение этой системы дает значения

для оценок параметров уравнения регрессии a и b :

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2}$$

$$a = \bar{y} - b \cdot \bar{x}$$

Решение уравнения регрессии в Excel

1. На листе Excel выделяют блок ячеек в котором

- строк всегда 5

- столбцов $-(m+1)$, где m – число независимых переменных

2. Вводят функцию: **ЛИНЕЙН(...)<Shift>+<Ctrl>+<Enter>**

Константа: =1, если параметр a присутствует в уравнении

=0, если уравнение имеет вид $y=b*x$

Статистика: =1, если необходима оценка достоверности

=0, если оценка не нужна

Решение уравнения регрессии в Excel

3. В выделенном блоке ячеек будет результат в виде

b_m	...	b_1	a
$\sigma(b_m)$		$\sigma(b_1)$	$\sigma(a)$
R^2	$\sigma(y)$		
$F_{\text{расч}}$	df		
$SS_{\text{регр.}}$	$SS_{\text{ост.}}$		

← значения параметров

← среднее квадр. отклонение полученных значений

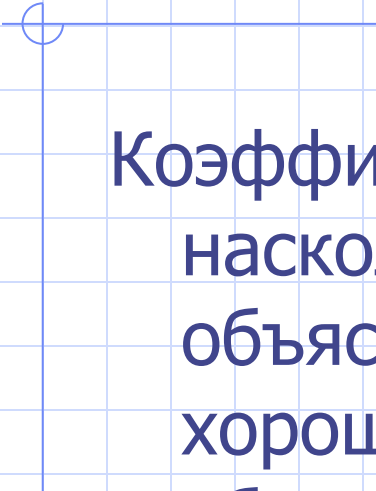
R^2 – коэффициент детерминации

$F_{\text{расч}}$ – расчетное значение функции Фишера

df – число степеней свободы ($=n-m-1$)

$SS_{\text{регр}}$ – регрессионная сумма квадратов

$SS_{\text{ост}}$ – остаточная сумма квадратов



Коэффициент детерминации показывает, насколько хорошо в выборке изменения Y объяснены изменениями X . Т. е., насколько хорошо выборочная модель регрессии объясняет поведение Y в выборке.

Изменения фактора Y измеряются его дисперсией $\sigma^2(Y)$.

Коэффициент детерминации:

$$R^2 = r_{yx}^2$$


- часть дисперсии Y , объясненная уравнением регрессии, т. е. изменениями в выборке фактора X .

Еще одно дополнение к R^2

Мы знаем, что $0 \leq R^2 \leq 1$.

Однако, если модель регрессии не имеет свободного члена, например, $Y = b \cdot x + e$, то возможны отрицательные значения R^2 .

Это также недостаток R^2 .



Статистические свойства
МНК-оценок коэффициентов
регрессии.

ТЕОРЕМА ГАУССА-МАРКОВА

Почему при оценке параметров модели а и b минимизируется именно

$$\sum_{i=1}^n e_i^2 \quad ?$$

Потому что при выполнении некоторых условий оценки а и b, полученные по МНК, оказываются очень хорошими:
несмещенными, эффективными, состоятельными.

Каких условий?

МНК-оценки a и b являются случайными величинами, свойства которых существенным образом зависят от свойств случайного члена e модели регрессии.

Условия Гаусса-Маркова

1. Математическое ожидание значений остатков e равно 0:

$$M(e_i) = 0 \text{ для всех наблюдений } x_i$$

2. Значение дисперсии ошибки является постоянной величиной $\sigma_{e_i}^2 = \sigma^2 = \text{const}$ для всех наблюдений X_i
(условие гомоскедастичности)

3. Значения e , для разных значений x_i независимы между собой

(отсутствие автокорреляции в остатках)

4. Значения x_i и e_i для одного и того же наблюдения независимы между собой $\sigma_{x_i, e_i} = 0$ для всех наблюдений

5. Модель является линейной относительно параметров

Условия Гаусса-Маркова

Для уравнения множественной регрессии:

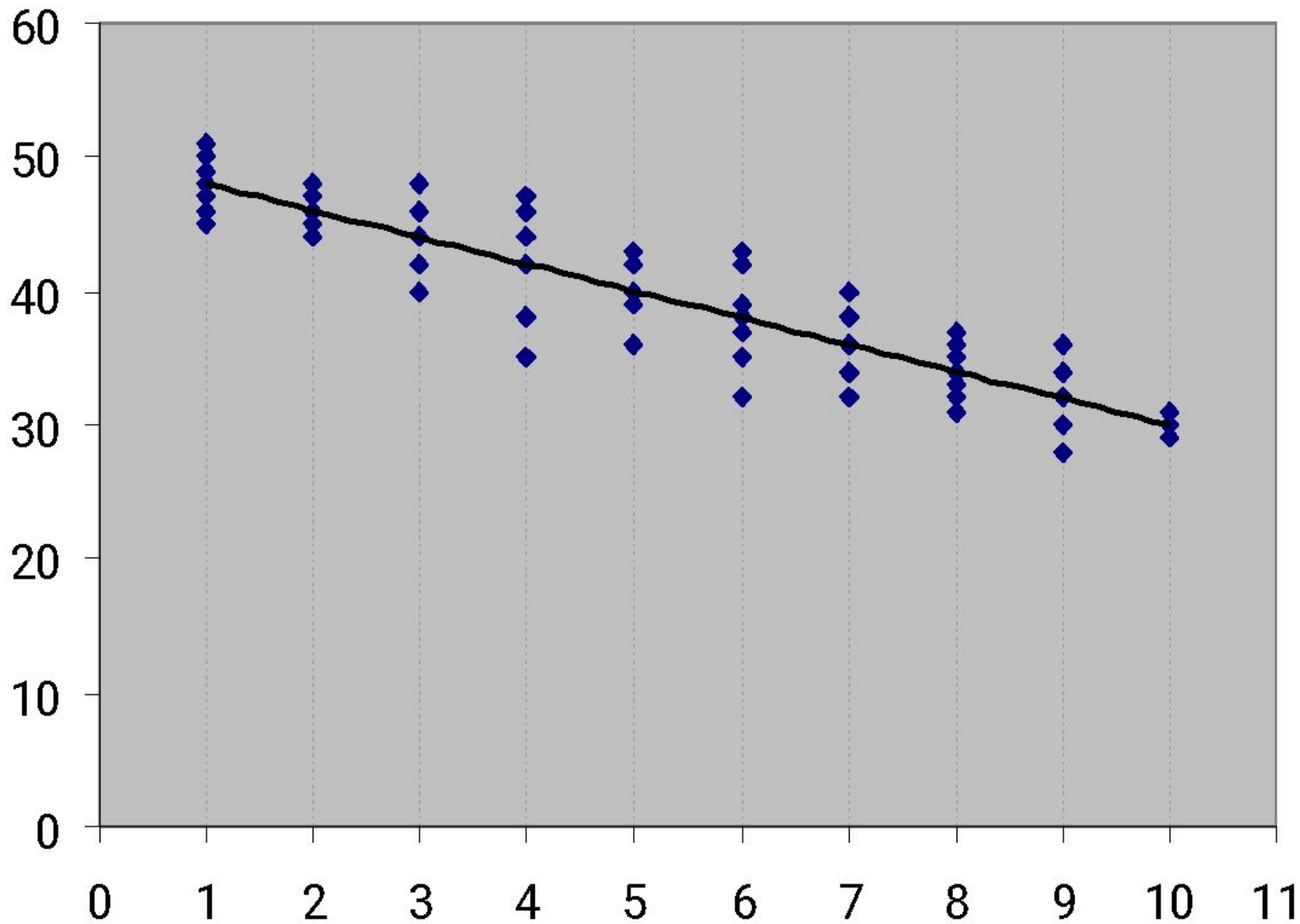
6. Факторы x_j независимы между собой в том смысле, что их выборочные парные линейные коэффициенты корреляции не превышают некоторого порога ρ :

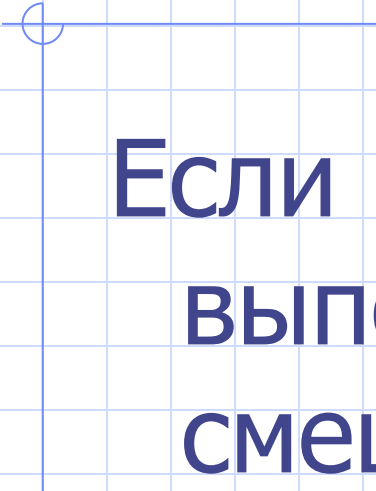
$$|r_{x_j x_i}| \leq \rho \quad (\text{условие отсутствия мультиколлинеарности})$$

7. Остатки являются нормально распределенной случайной величиной, т.е. подчиняются закону нормального распределения.

Модель, удовлетворяющая предпосылкам МНК (1)-(7), называется классической нормальной моделью регрессии,

если не выполняется только условие (7), то модель – классическая модель регрессии.



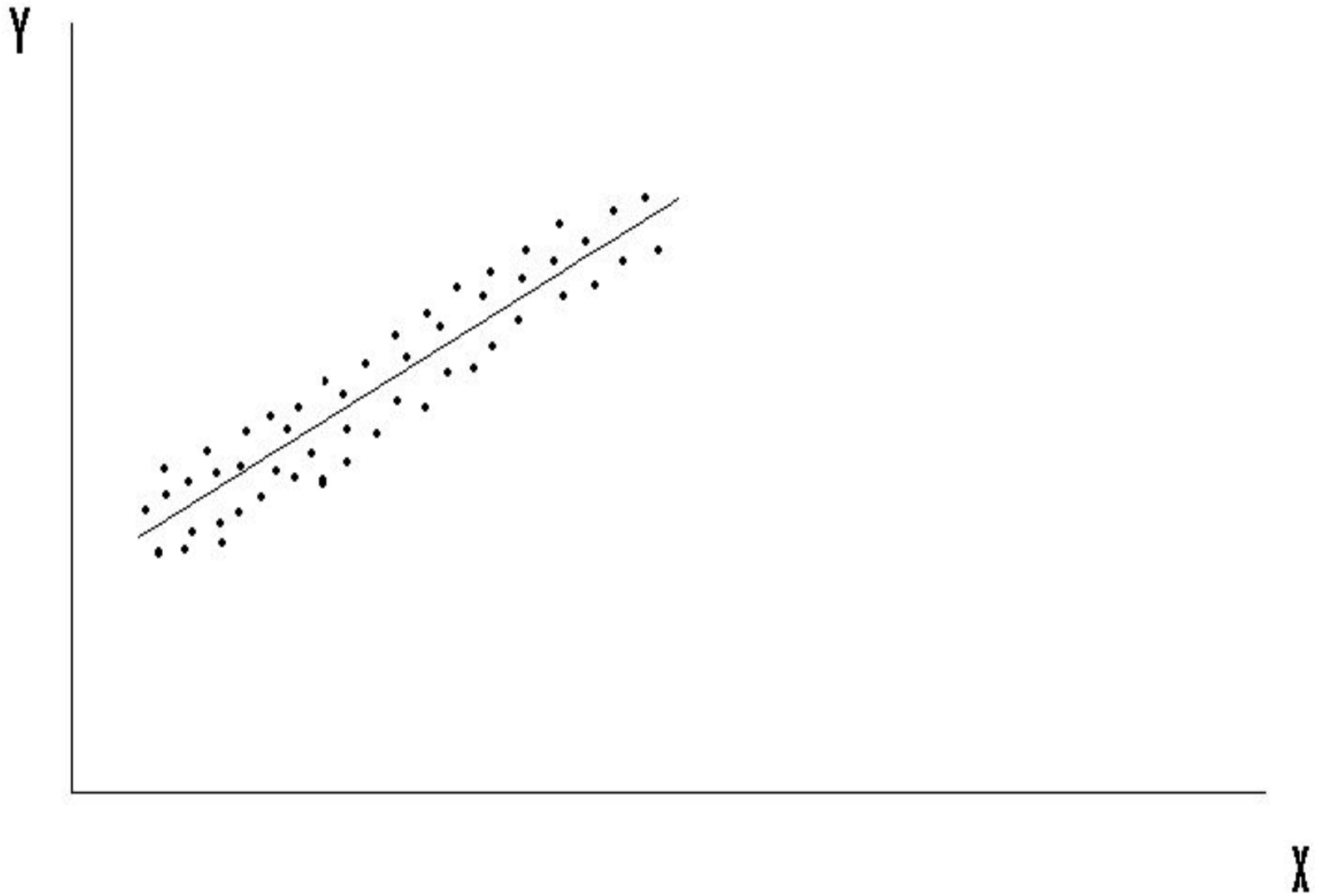


Если 1-е условие Г-М не выполняется, МНК дает смещенную оценку для b .

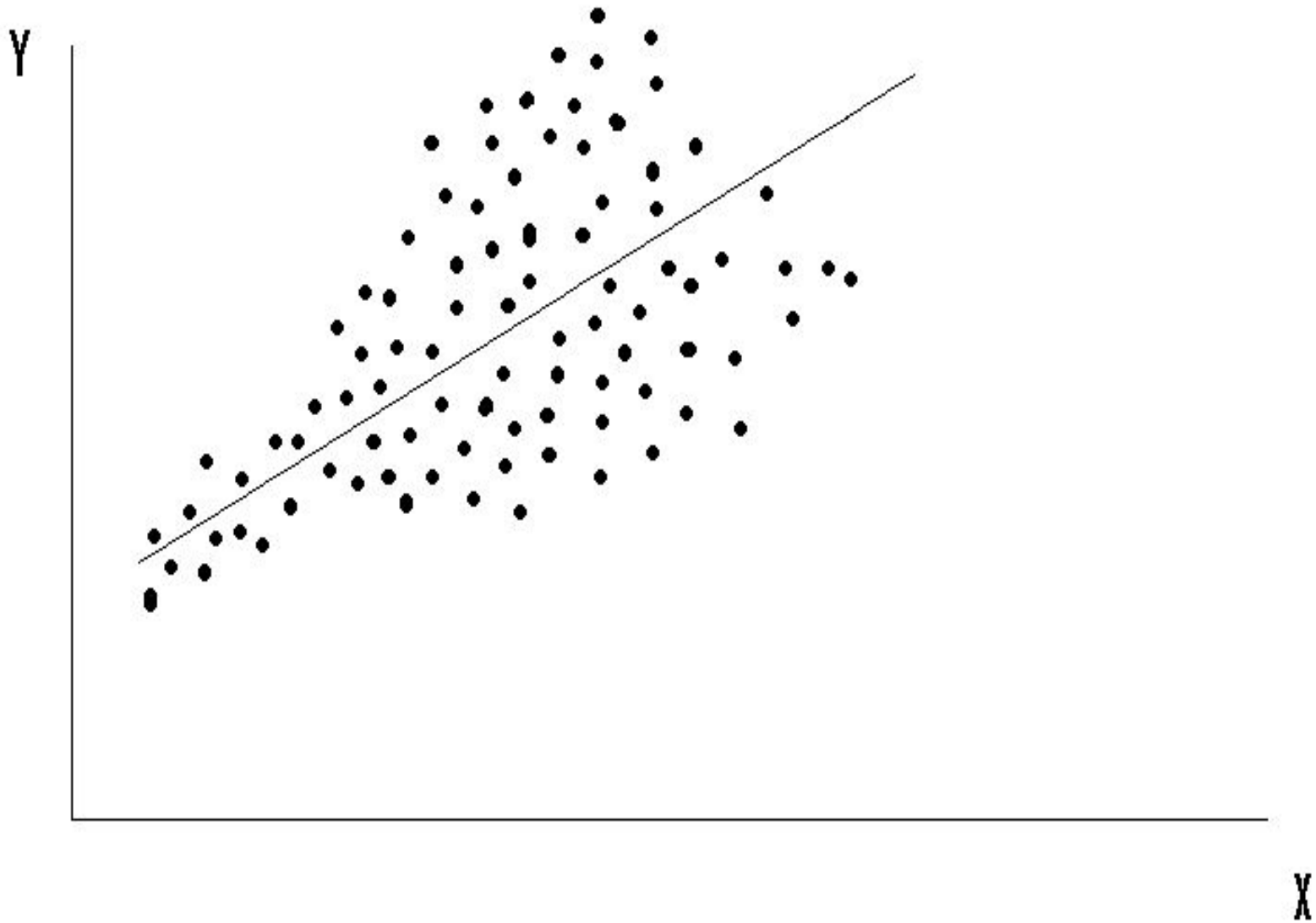
2. $\sigma_{ei}^2 = \sigma^2 = \text{const}$ для всех наблюдений X_i

Условие гомоскедастичности ошибок.

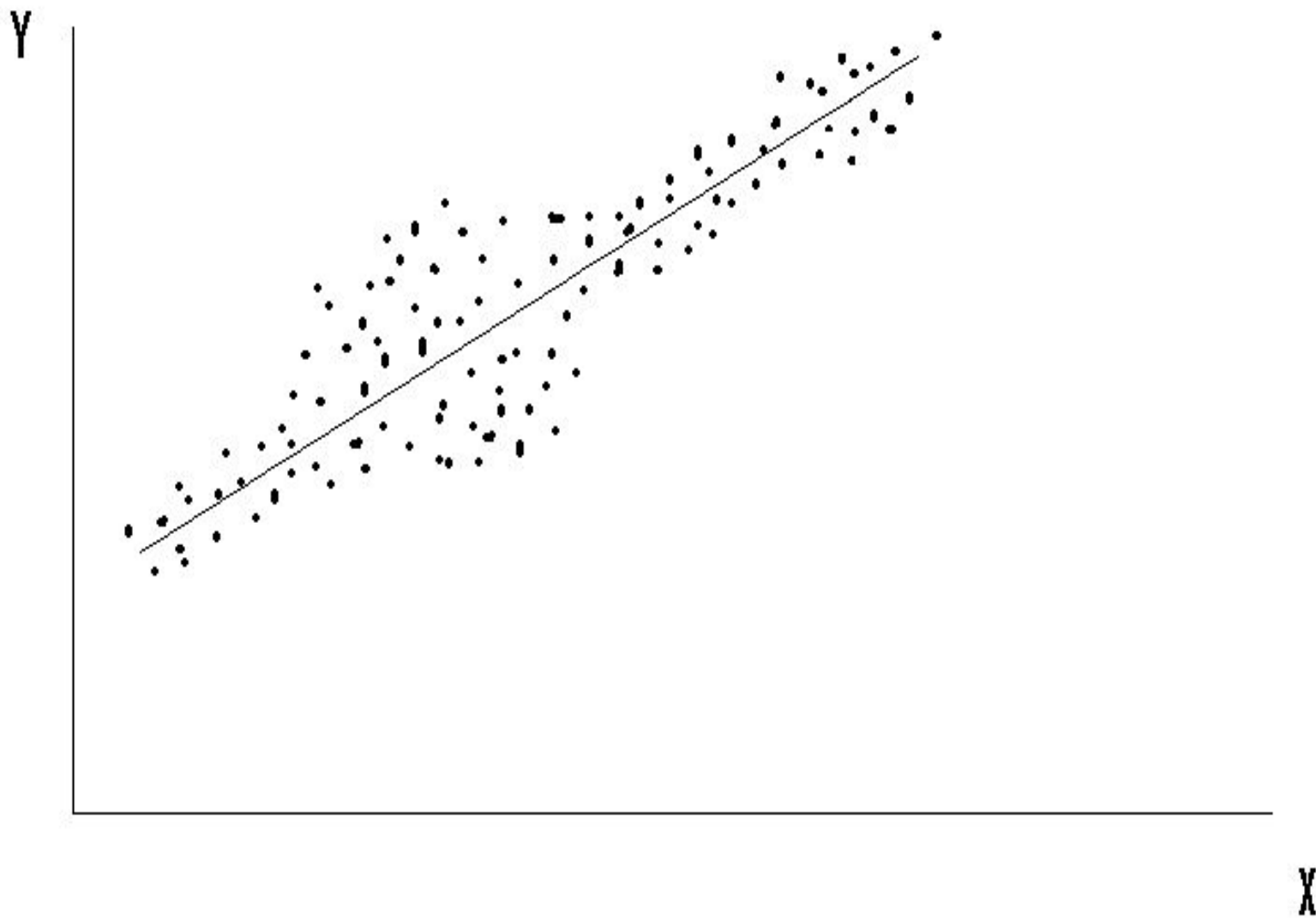
Когда оно не выполняется, говорят о гетероскедастичности ошибок.



2-е условие Г-М выполняется.



2-е условие Г-М не выполняется.



2-е условие Г-М не выполняется.

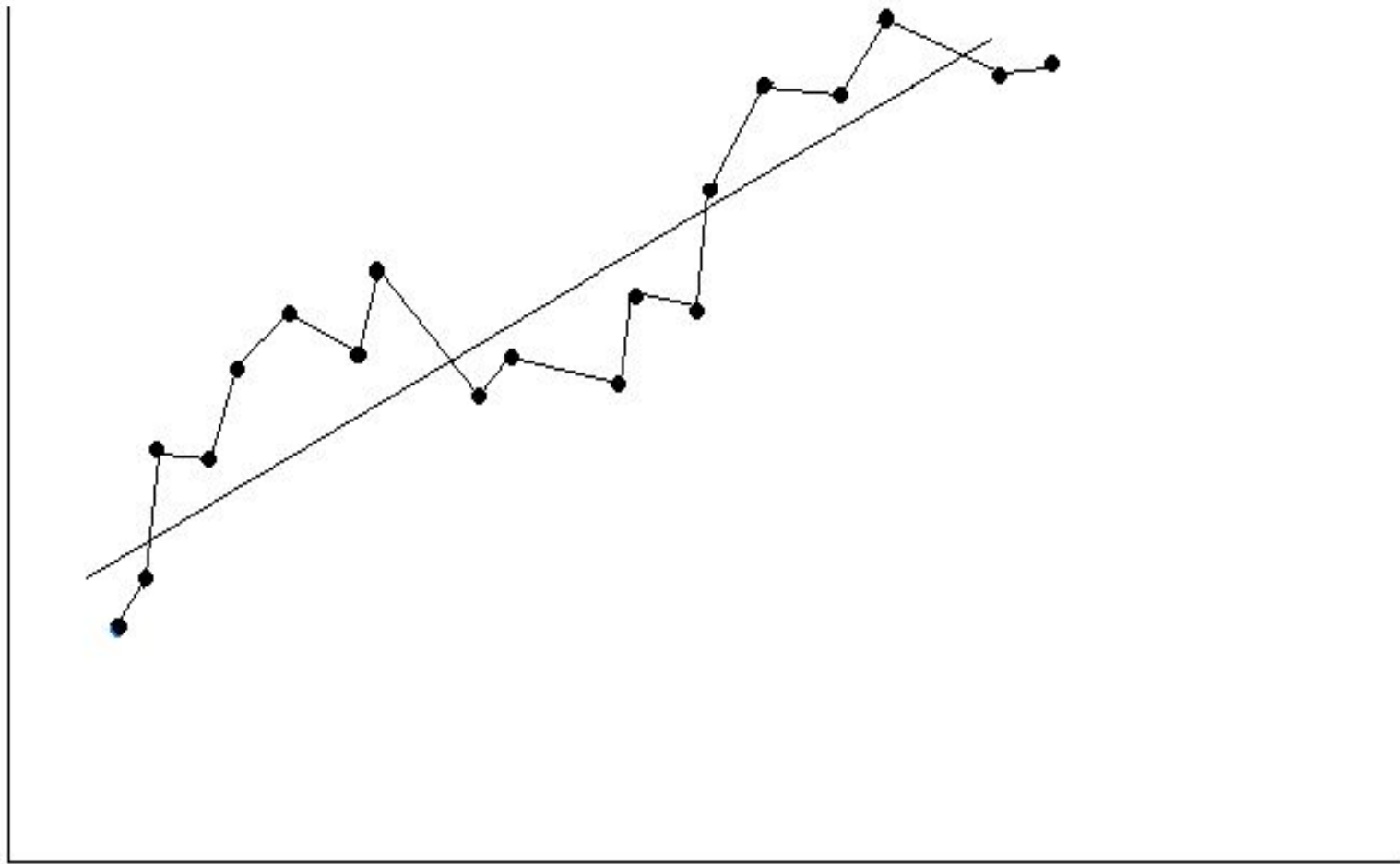
3. $\sigma_{e_i, e_j} = 0$ для всех X_i и X_j , $i \neq j$.

Условие некоррелированности ошибок для разных наблюдений.

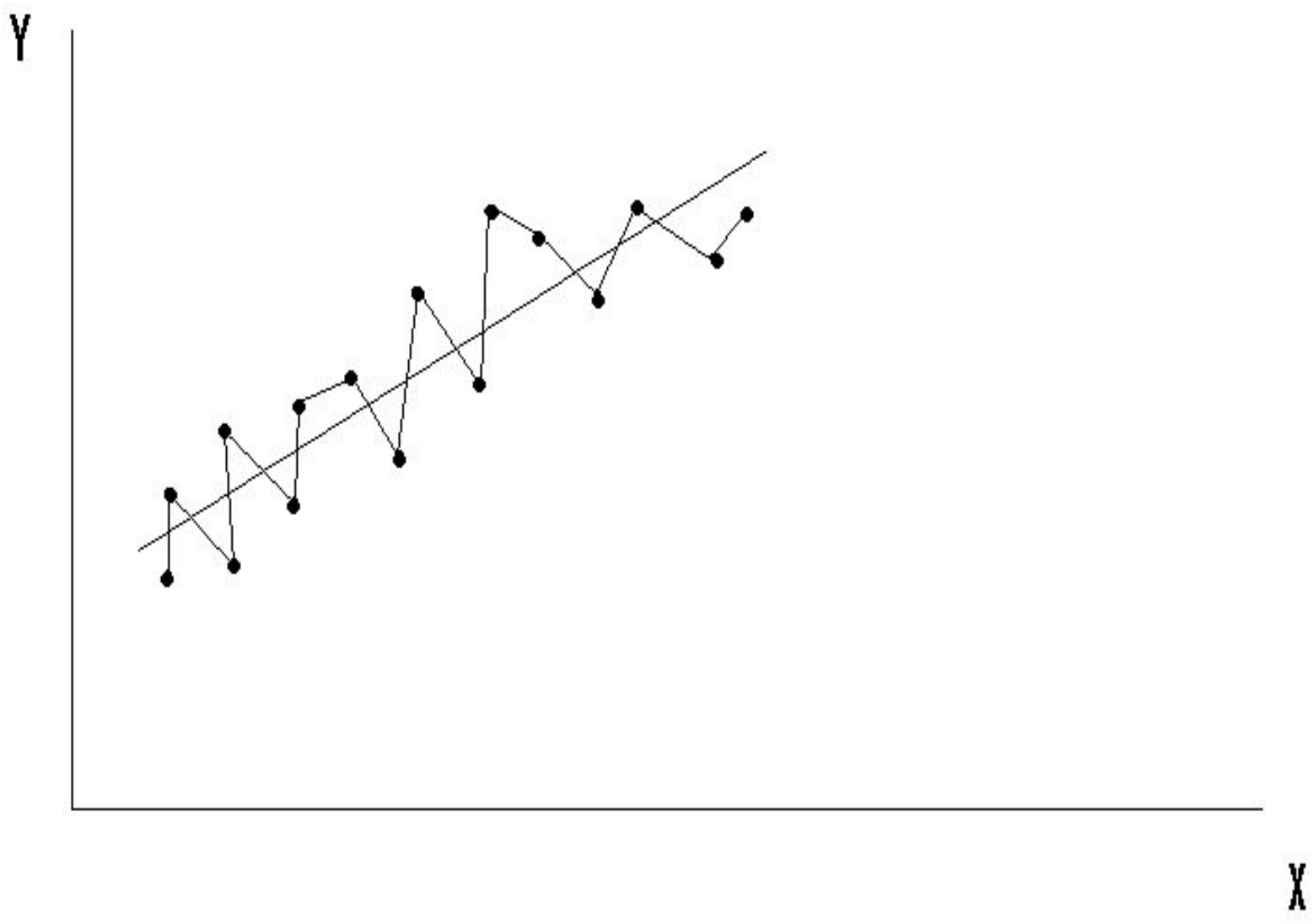
Это условие часто нарушается, когда данные являются временными рядами, из-за наличия в динамике экономических показателей различных регулярных колебаний.

При невыполнении (3) говорят об автокоррелированности остатков.

Y



X



4. $\sigma_{X_i, e_i} = 0$ для всех наблюдений.

Случайный член распределен независимо от объясняющей переменной.

Это всегда выполняется, если объясняющие переменные не являются случайными величинами.

Дополнительное условие:

7. Случайный член, $i=1, \dots, n$,
имеет нормальное
распределение,
$$e_i \sim N(0, \sigma_e^2)$$

Это условие не нужно для обеспечения хороших свойств оценок a и b .

Но оно позволяет корректно проводить проверку гипотез о коэффициентах регрессии.

Реальность предположения о нормальности e_i обеспечивается Центральной предельной теоремой.

Теорема Гаусса-Маркова

Если предпосылки МНК соблюдаются, то оценки, полученные по МНК, обладают следующими свойствами:

1. Оценки параметров являются **несмещенными**, т.е. $M(b_i) = b_i$ и $M(a) = a$. Это вытекает из того, что $M(e_i) = 0$ и говорит об отсутствии систематической ошибки в определении положения линии регрессии
2. Оценки параметров **состоятельны**, т.к. дисперсия оценок параметров при возрастании числа n наблюдений стремится к нулю. Т.е. При увеличении объема выборки надежность оценок возрастает.
3. Оценки параметров **эффективны**, т.е. Они имеют наименьшую дисперсию по сравнению с другими оценками данных параметров.

F-критерий Фишера

H_0 – гипотеза о статистической незначимости уравнения регрессии и показателя тесноты связи: $b=0, r_{yx}=0$

$$F = \frac{R^2}{1 - R^2} (n - m - 1)$$

Если $F_{расч} \geq F_{табл}$, то отвергается гипотеза H_0 и признается значимость и надежность полученных оценок параметров a и b

t-статистика Стьюдента

H_0 – гипотеза о статистической незначимости оценок параметров уравнения регрессии и показателя тесноты связи: $a=b=r_{yx}=0$

$$t_b = \frac{b}{m_b}; \quad t_a = \frac{a}{m_a}; \quad t_r = \frac{r}{m_r}$$

где m_b, m_a, m_r – случайные ошибки параметров линейной регрессии и коэффициента корреляции

Формулы для расчета случайных ошибок:

$$m_b = \sqrt{\frac{\sum (y - \hat{y})^2}{(n - m - 1) \cdot \sum (x - \bar{x})^2}}$$

$$m_a = \sqrt{\frac{\sum (y - \hat{y})^2}{(n - m - 1)} \cdot \frac{\sum x^2}{n \cdot \sum (x - \bar{x})^2}}$$

$$m_r = \sqrt{\frac{1 - r_{yx}^2}{n - m - 1}}$$

Расчет доверительного интервала прогноза

$$\gamma_a = a \pm \Delta_a$$

$$\gamma_b = b \pm \Delta_b$$

где

$$\Delta_a = t_{табл} m_a, \quad \Delta_b = t_{табл} m_b$$

Если в границы доверительного интервала попадает ноль, т.е. нижняя граница отрицательна, а верхняя положительна, то оцениваемый параметр принимается нулевым, т.к. он не может одновременно принимать и положительное и отрицательное значение.

Расчет прогнозного значения

Прогнозное значение y_p определяется путем подстановки в уравнение регрессии соответствующего (прогнозного) значения x_p .
Вычисляется средняя стандартная ошибка прогноза:

$$m_{\hat{y}_p} = \sqrt{\frac{\sum (y - \hat{y})^2}{n - m - 1}} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

и строится доверительный интервал прогноза:

$$\gamma_{\hat{y}_p} = \hat{y}_p \pm t_{табл} \cdot m_{\hat{y}_p}$$