
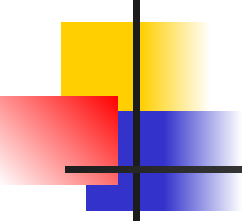


Лекция 1

Линейная парная регрессия

- 
-
- 1. Эконометрика как научная дисциплина.*
 - 2. Спецификация уравнения парной регрессии.*
 - 3. Параметризация уравнения парной регрессии.*
 - 4. Экономическая интерпретация параметров модели.*

1. Эконометрика как научная дисциплина.



Эконометрика – это дисциплина, изучающая методы построению на основе статистических данных аналитических математических моделей экономических процессов с целью их анализа и прогнозирования.

Основными целями эконометрики являются:

- объяснение поведения экономических и социальных показателей;
- прогнозирование показателей;
- управление показателями.



Основными задачами эконометрики считаются:

- выбор типа эконометрических моделей (регрессионные модели с одним уравнением, системы взаимозависимых уравнений, модели временных рядов);
- оценка параметров выбранных моделей;
- проверка качества модели и ее полученных параметров;
- практическое использование построенных адекватных моделей для анализа, прогноза и осмысленного управления экономической политикой.

Для решения поставленных задач широко используется инструментарий, основу которого составляют основные положения теории вероятностей и математической статистики.

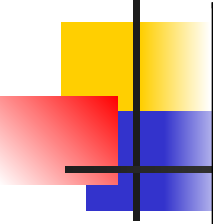


2. Спецификация уравнения парной регрессии.

Изучение зависимостей экономических переменных начнём со случая двух переменных. Обозначим их символами x и y .

Первую из них (x) будем называть привычным термином *независимая переменная*, или, как принято в эконометрике, - *объясняющая* переменная или *фактор*. Вторую (y) – *зависимой* или *объясняемой* переменной.

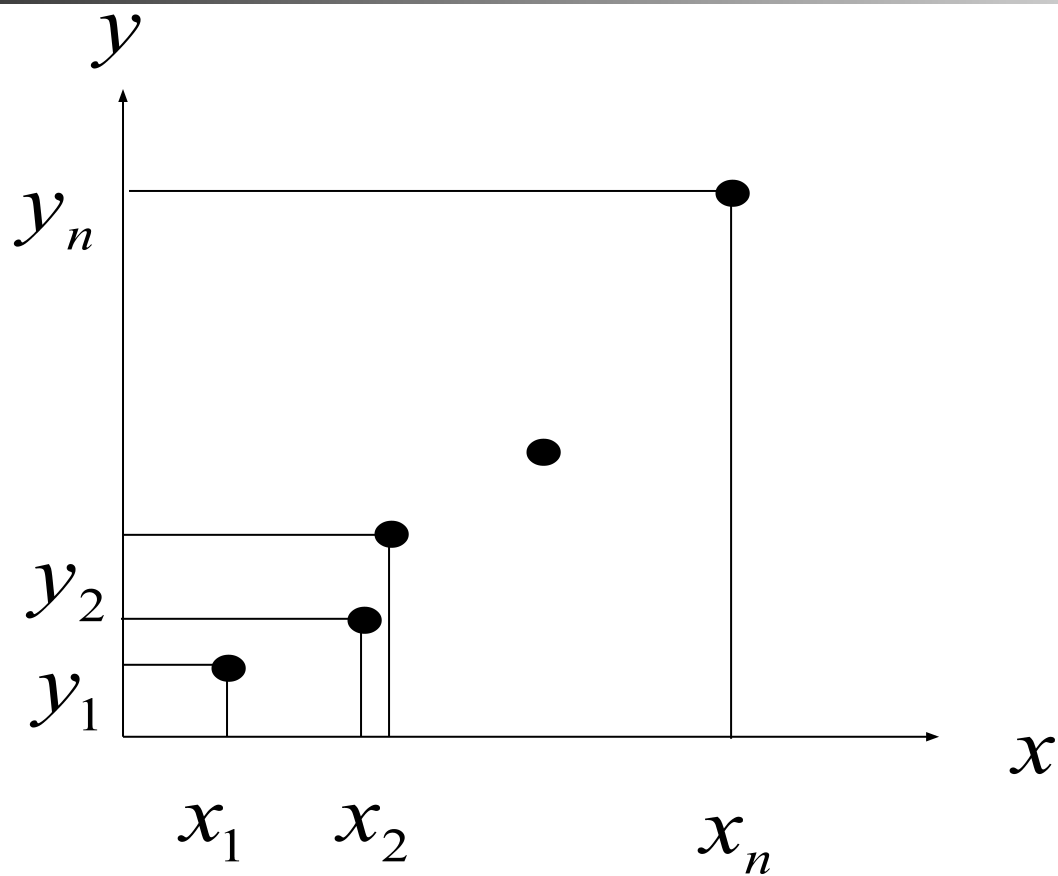
Предположим, что имеются статистические данные в виде рядов значений указанных переменных, которые сведены в следующую таблицу (двумерная выборка объема n , $n > 7$)

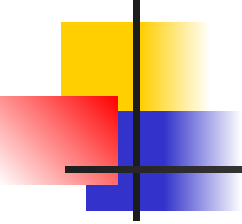


y	y_1	y_2	y_3	\dots	y_n
x	x_1	x_2	x_3	\dots	x_n

Если нанести соответствующие точки $(x_i, y_i), i = \overline{1, n}$ на координатную плоскость, то получим график, который называют *полем корреляции* или *диаграммой рассеивания*.

Рис. 1. Поле корреляции





Построенные точки на практике никогда не будут располагаться на некоторой *гладкой* линии типа прямой, параболы, экспоненты и т.д.

Происходит это потому, что на зависимую переменную помимо фактора влияют другие, либо неучтенные, либо случайные факторы.

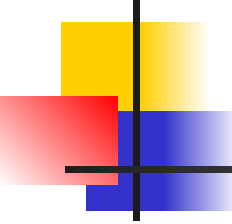
Связь переменных, на которую накладывается воздействие случайных факторов, называется *статистической* связью.

Формула статистической связи между переменными

$$y = f(x, \varepsilon) ,$$

где ε — случайный фактор, называется *уравнением регрессии*.

Для двух переменных она называется уравнением *парной* регрессии.



Выбор конкретной формулы связи для двух переменных называется *спецификацией* уравнения регрессии.

Класс математических функций для описания связи двух переменных очень широк:

линейная функция

$$y = \beta_0 + \beta_1 x + \varepsilon \quad ;$$

парабола второго порядка

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \quad ;$$

степенная функция

$$y = \beta_0 x^{\beta_1} + \varepsilon$$

и т. д.

Построенное поле корреляции иногда помогает
выполнить спецификацию уравнения регрессии
(рис.2, рис. 3).

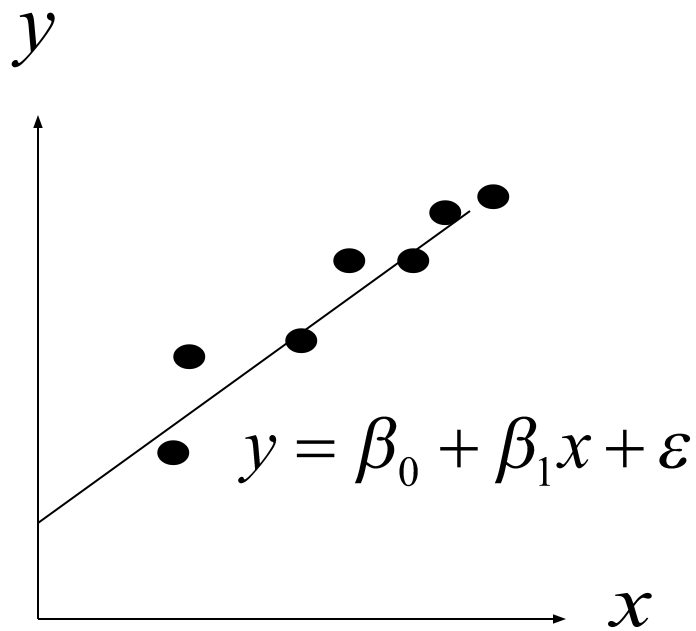


Рис. 2

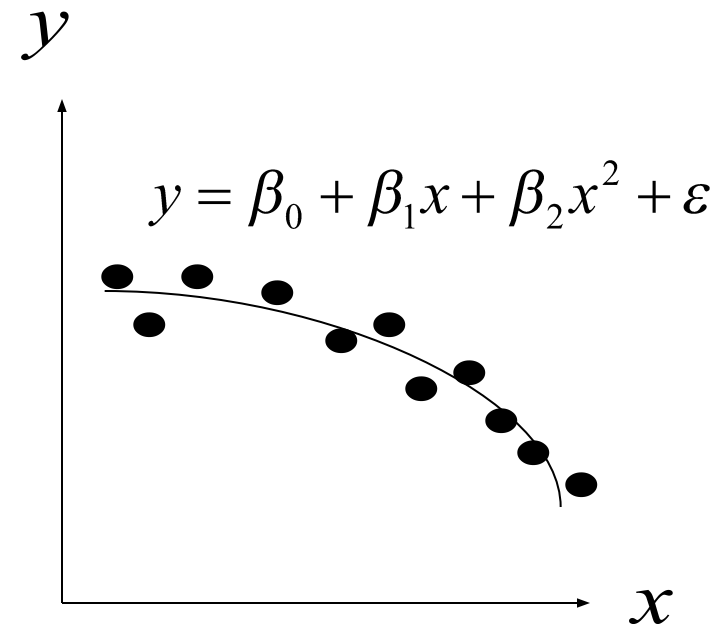
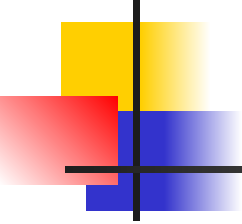


Рис. 3



В других случаях поле корреляции не позволяет сказать что-то определенное о виде зависимости между x и y (рис. 4).

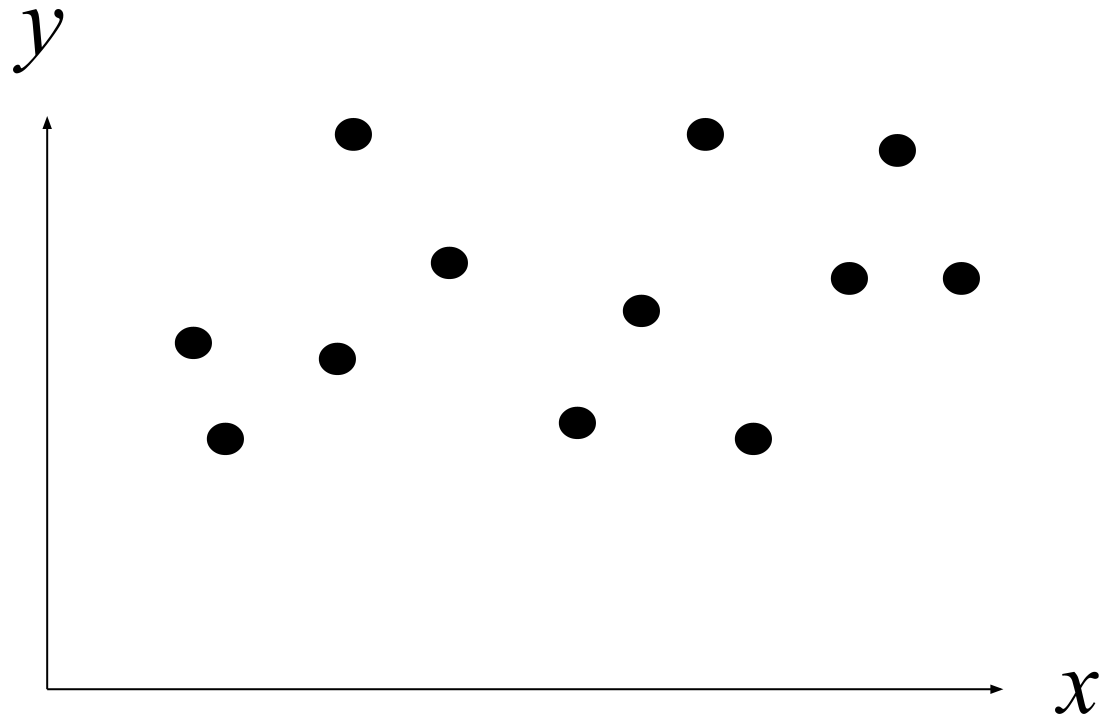
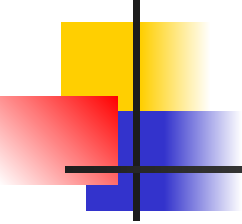


Рис. 4



В последнем случае начинают с выбора наиболее простой связи – *линейной* :

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Если в последующем она не устраивает по точности, то ищут какую-либо нелинейную связь между переменными x и y .

3. Параметризация уравнения парной регрессии.

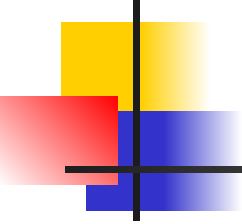
После выбора формы связи между переменными выполняется оценка значений коэффициентов выбранной модели с использованием имеющихся статистических данных.

Этот процесс называется *параметризацией* уравнения регрессии.

Пусть поле корреляции имеет вид как на рис. 2.

Иначе между переменными x и y в природе существует некоторая линейная зависимость и статистические данные должны удовлетворять уравнению:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = \overline{1, n} . \quad (1)$$



Здесь β_0, β_1 – коэффициенты, которые подлежат определению, ε_i – случайная составляющая в i -м наблюдении, которую называют ***возмущением***.

Уравнение (1) называют ***модельным*** уравнением регрессии.

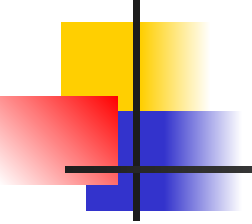


Возникает задача:

по выборке $x_i, y_i, i = \overline{1, n}$ оценить уравнение (1) и такой оценкой является **выборочное** уравнение регрессии:

$$\tilde{y} = b_0 + b_1 x \quad (2)$$

Построение уравнения (2) сводится к получению точечных оценок b_0, b_1 неизвестных коэффициентов β_0, β_1 модельного уравнения (1). Классический подход к оцениванию неизвестных коэффициентов основан на методе наименьших квадратов (МНК).



Чтобы при этом точечные оценки b_0, b_1 были "хорошими" (несмещенными, эффективными и состоятельными) требуется сделать следующие предположения относительно уравнения (1).

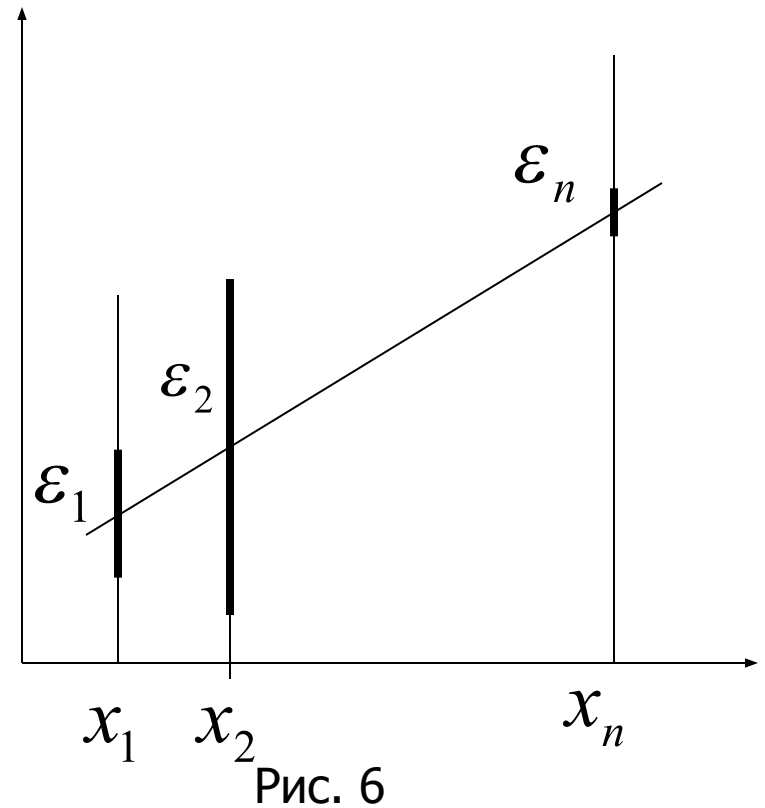
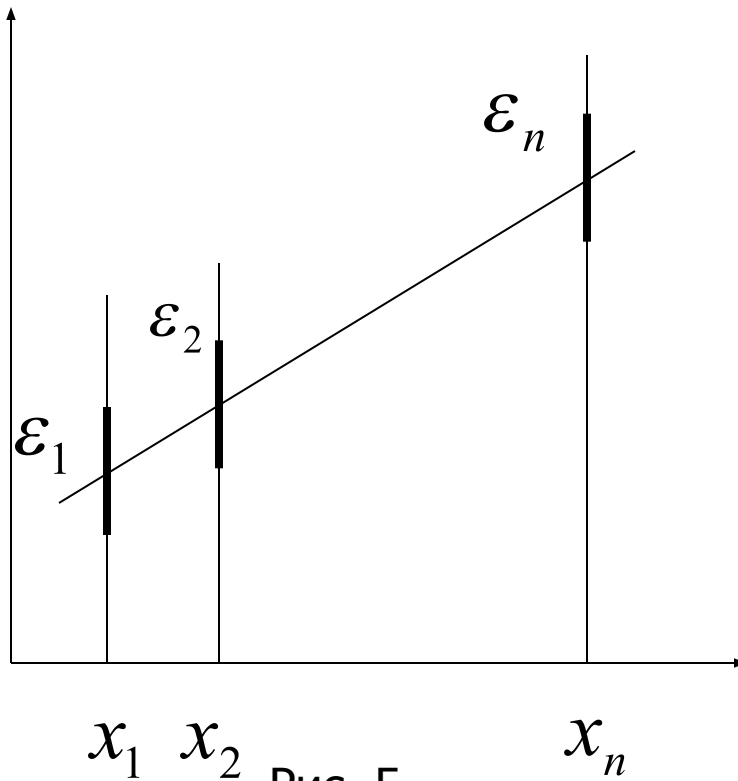
1°. Возмущения ε_i являются случайными величинами, а переменные x_i таковыми не являются.

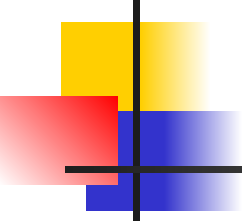
2°. Математическое ожидание случайных величин ε_i во всех наблюдениях равно нулю, т. е.

$$M(\varepsilon_i) = 0 \quad i = \overline{1, n}$$

3°. Дисперсия случайных величин ε_i постоянна для любого наблюдения:

$$D(\varepsilon_i) = \sigma^2 = \text{const} < \infty \quad .$$





Это свойство называют *гомоскедастичностью* возмущений (одинаковый разброс, рис.5).

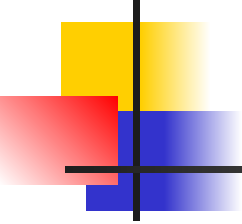
Если оно не выполняется, то говорят о *гетероскедастичности* (различный разброс, рис. 6).

4°. Возмущения ε_i и ε_j в различных наблюдениях являются некоррелированными величинами

$$M(\varepsilon_i \cdot \varepsilon_j) = 0.$$

5°. Случайные величины ε_i имеют нормальный закон распределения

$$\varepsilon_i \sim N(0, \sigma^2) \quad .$$



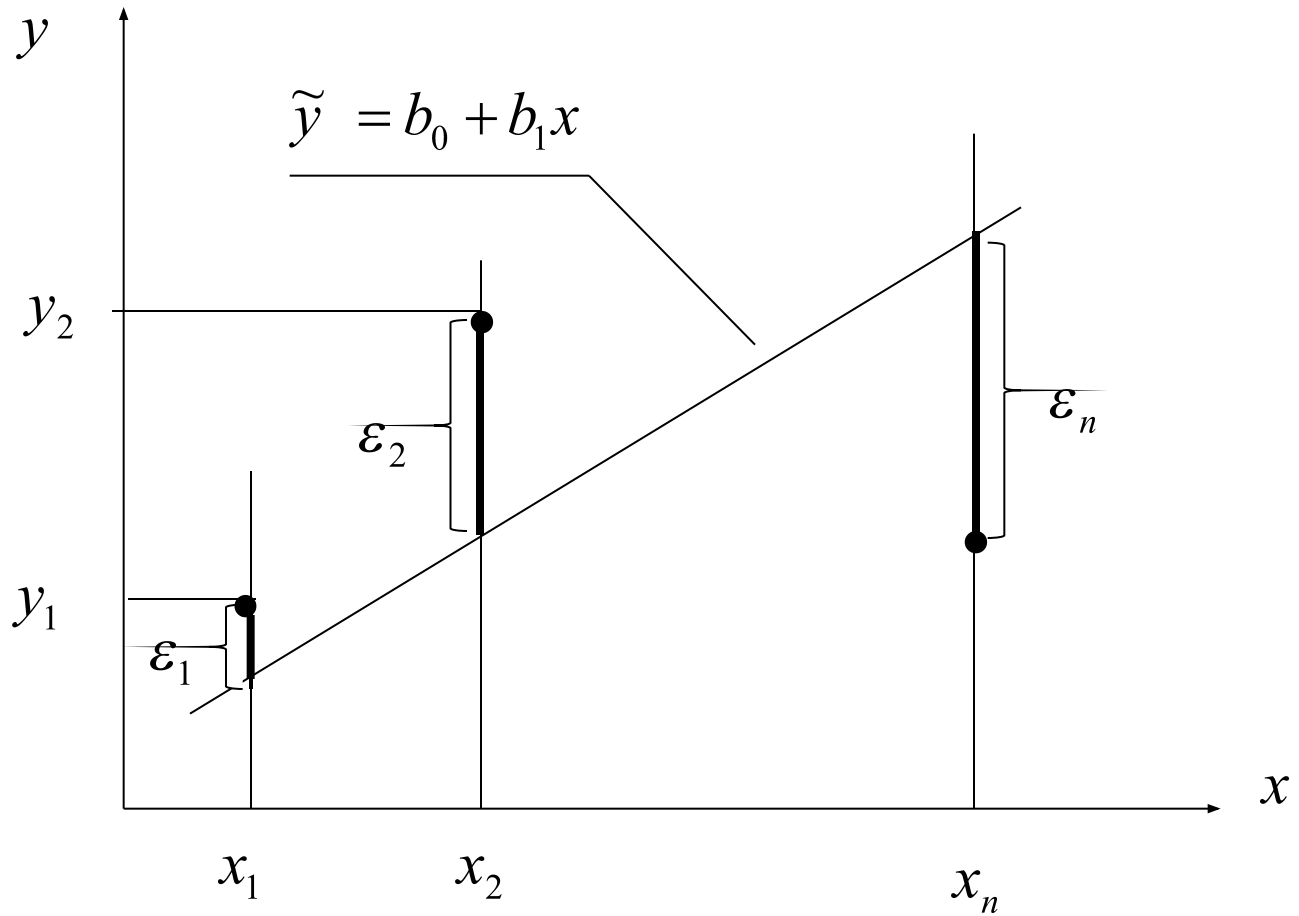
При выполнении указанных требований, которые называют *предпосылками МНК*, модель (1) называют *нормальной классической линейной регрессионной моделью*.

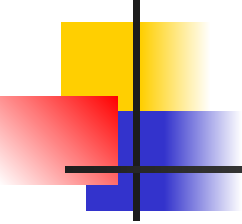
Согласно МНК параметры b_0, b_1 определяются таким образом, чтобы сумма квадратов отклонений выборочных значений y_i от значений \tilde{y}_i , полученных по уравнению регрессии

$$\tilde{y}_i = b_0 + b_1 x_i$$

по всем $i = \overline{1, n}$ наблюдениям была минимальной.

Рис. 7





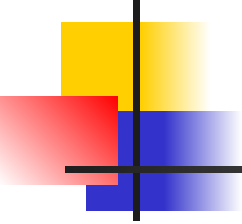
Если ввести в рассмотрение величины $e_i = y_i - \tilde{y}_i$, называемые *остатками*, то параметры b_0, b_1 находят из условия минимума функции двух переменных

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

Применяя необходимые условия экстремума функции двух переменных

$$\frac{\partial Q}{\partial b_0} = 0 \quad , \quad \frac{\partial Q}{\partial b_1} = 0 \quad ,$$

после несложных преобразований можно получить так называемую *систему нормальных уравнений*


$$\begin{cases} b_0 n + b_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i, \\ b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \end{cases} \quad (4)$$

для определения двух неизвестных b_0 и b_1 .

Разделив обе части уравнений на n , получим систему в виде

$$\begin{cases} b_0 + b_1 \bar{x} = \bar{y}, \\ b_0 \bar{x} + b_1 \overline{x^2} = \overline{xy}, \end{cases}$$

где

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i,$$

$$\overline{x^2} = \frac{1}{n} \sum_{i=1}^n x_i^2, \quad \overline{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i.$$

Решая последнюю систему, получим

$$b_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{x^2 - \bar{x}^2},$$

(5)

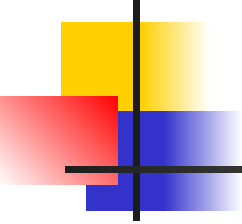
$$b_0 = \bar{y} - b_1 \bar{x}.$$

Оценки (5) называют **МНК-оценками**.

Теорема Гаусса-Маркова.

Если регрессионная модель (1) удовлетворяет предпосылкам 1-4, то оценки (5) имеют наименьшую дисперсию в классе всех несмещенных оценок.

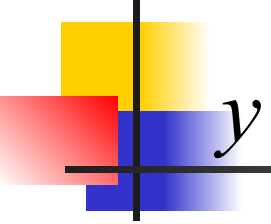
Другими словами они являются **эффективными** и **несмещенными**.



4. Экономическая интерпретация параметров модели.

Параметр b_1 в уравнении регрессии (2) называют *выборочным коэффициентом регрессии* y по x . Он показывает, на сколько единиц в среднем изменится переменная y при увеличении переменной x на одну единицу своего измерения. В этом экономический смысл параметра b_1 .

Параметр b_0 в общем случае не имеет экономического смысла. Формально - это значение переменной y при $x = 0$, если такое возможно.



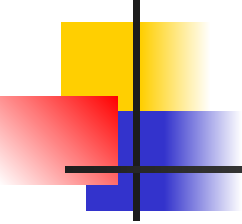
Из экономического смысла b_1 следует, что он является измерителем *тесноты связи* переменных y и x . Однако его значение существенно зависит от единиц измерения данных переменных.

Например, если y измеряется не в тоннах, а в кг, то x уменьшается в 1000 раз.

Характеристикой тесноты связи, не зависящей от единиц измерения, является величина

$$r_{xy} = b_1 \frac{\sigma_x}{\sigma_y}, \quad (6)$$

где \bar{y} и \bar{x} — выборочные средние, σ_y и σ_x — квадратические отклонения переменных y и x соответственно.



Величину r_{xy} называют **выборочным коэффициентом парной корреляции** и он показывает, на сколько средних квадратических

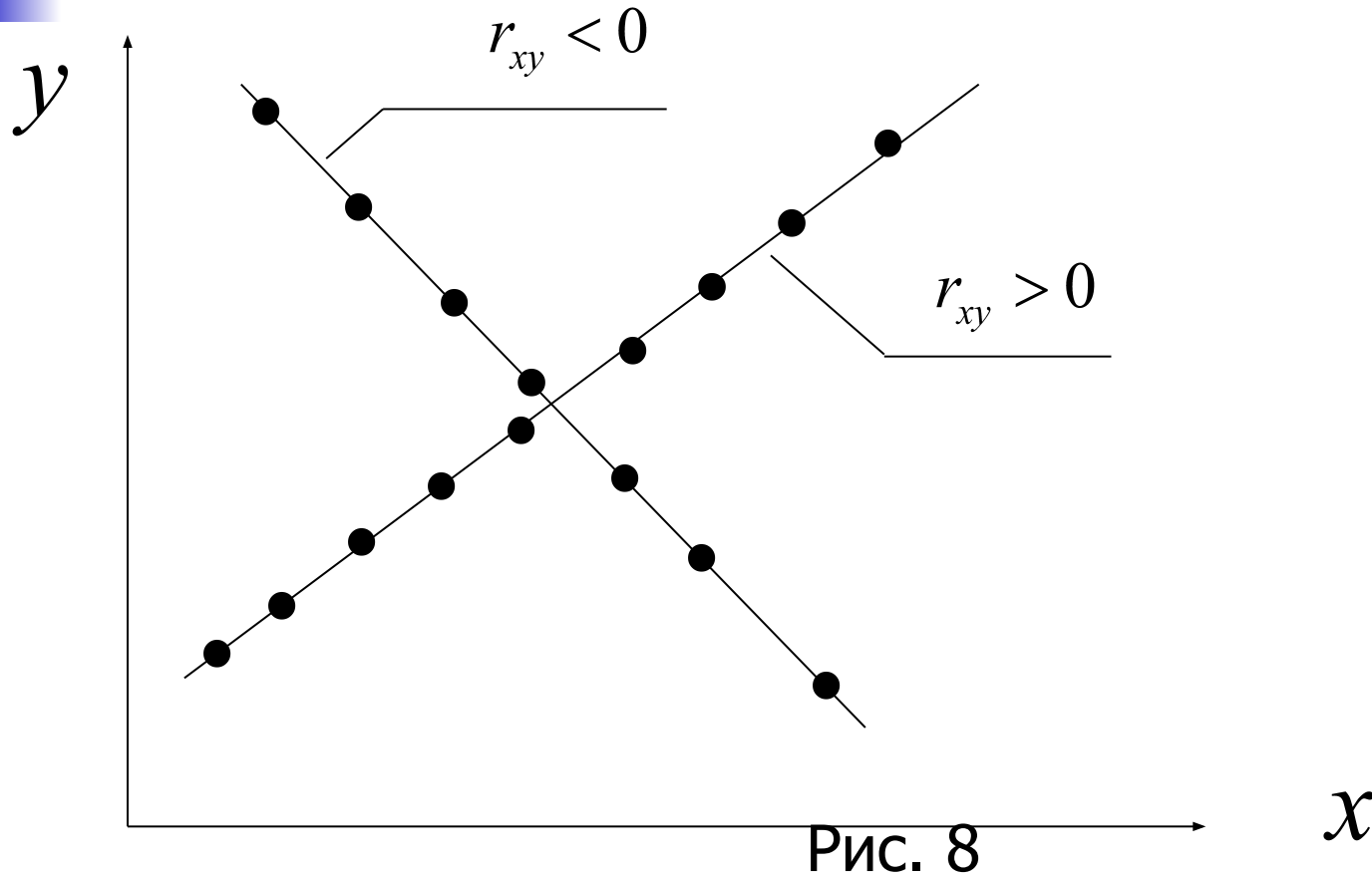
отклонений σ_y изменяется в среднем переменная y , когда x увеличится на одно σ_x .

Коэффициент r_{xy} обладает следующими свойствами.

1. Значения принадлежат отрезку: $[-1, 1]$. Чем ближе $|r_{xy}|$ к 1, тем теснее связь переменных y и x .

Если $r_{xy} > 0$, то эта связь **прямая**, в противном случае, когда $r_{xy} < 0$ - **обратная**.

2. При $r_{xy} = \pm 1$ связь представляет линейную функциональную зависимость между переменными y и x (рис. 8).



3. Если $r_{xy} = 0$, то корреляционная связь между переменными y и x отсутствует и $\tilde{y} = \bar{y}$ (рис. 9).

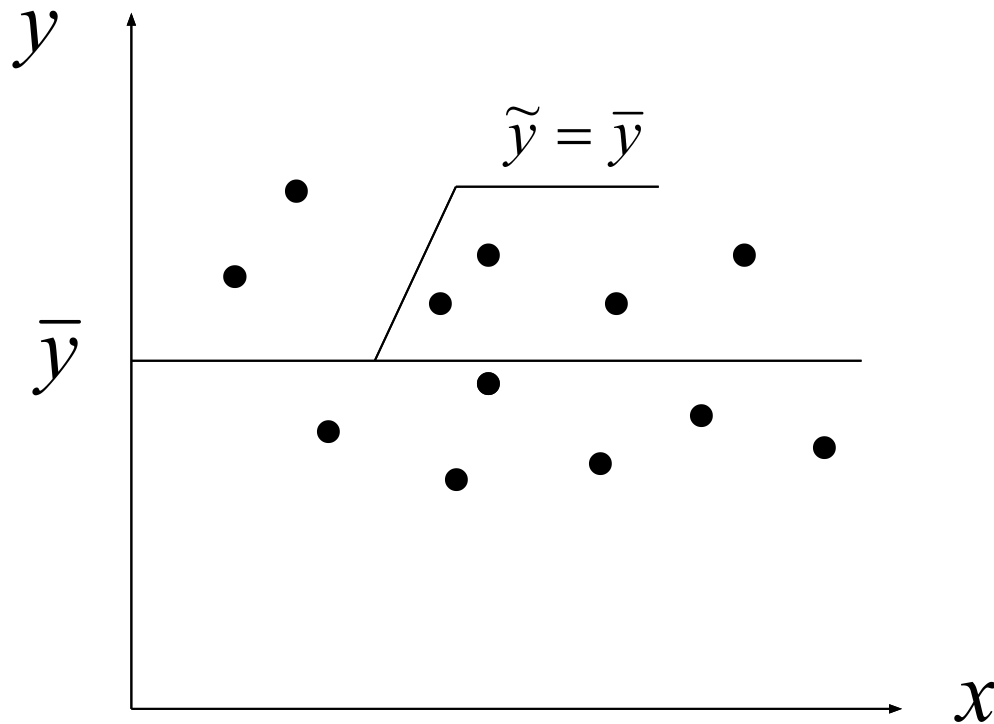
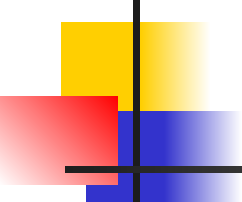


Рис. 9

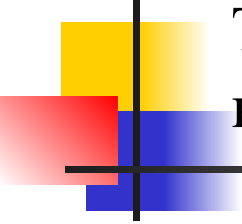


При других значениях $|r_{xy}|$ характеристику тесноты связи даёт *шкала Чеддока*:

$ r_{xy} $	0,1-0,3	0,3-0,5	0,5-0,7	0,7-0,9	0,9-0,99
Характеристика связи	слабая	умеренная	заметная	высокая	весьма высокая

Другим важным показателем силы связи фактора с результатом является *коэффициент эластичности*.

Различают *средние* и *точечные* коэффициенты эластичности.



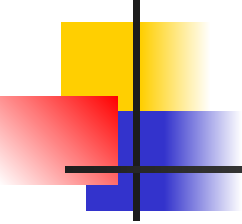
Пусть уравнение регрессии имеет вид $\tilde{y} = f(x)$.
Тогда средний коэффициент эластичности находится из
выражения

$$\bar{\varepsilon} = \frac{df}{dx} \cdot \frac{\bar{x}}{f(\bar{x})}. \quad (7)$$

В частности, для линейной зависимости верна формула

$$\bar{\varepsilon} = b_1 \cdot \frac{\bar{x}}{\bar{y}}. \quad (8)$$

Коэффициент $\bar{\varepsilon}$ является *безразмерным*, и он показывает, на сколько процентов изменится переменная y относительно своего среднего уровня при росте фактора x на 1% от среднего значения.



Точечный коэффициент эластичности
рассчитывается для конкретного значения
переменной $x = x_0$ по формуле

$$\mathcal{E}_0 = \frac{df(x_0)}{dx} \cdot \frac{x_0}{f(x_0)} ,$$

в частности, для линейной модели $\tilde{y} = b_0 + b_1x$
она запишется

$$\mathcal{E}_0 = b_1 \cdot \frac{x_0}{b_0 + b_1x_0} .$$