Парная линейная регрессия и метод наименьших квадратов

Лекция

Цели лекции

- Раскрыть понятие регрессии.
- Познакомиться с методом наименьших квадратов – методом построения линейного уравнения регрессии.

Виды зависимостей между переменными

1. Функциональные: $Y = f(X)$.

Имеют место при исследовании связей между неслучайными переменными. Такие связи в эконометрике не рассматриваются.

2. Статистические: изменение одной из величин влечет изменение закона распределения другой (доход – потребление, цена – спрос и т.д.).

Виды статистических зависимостей

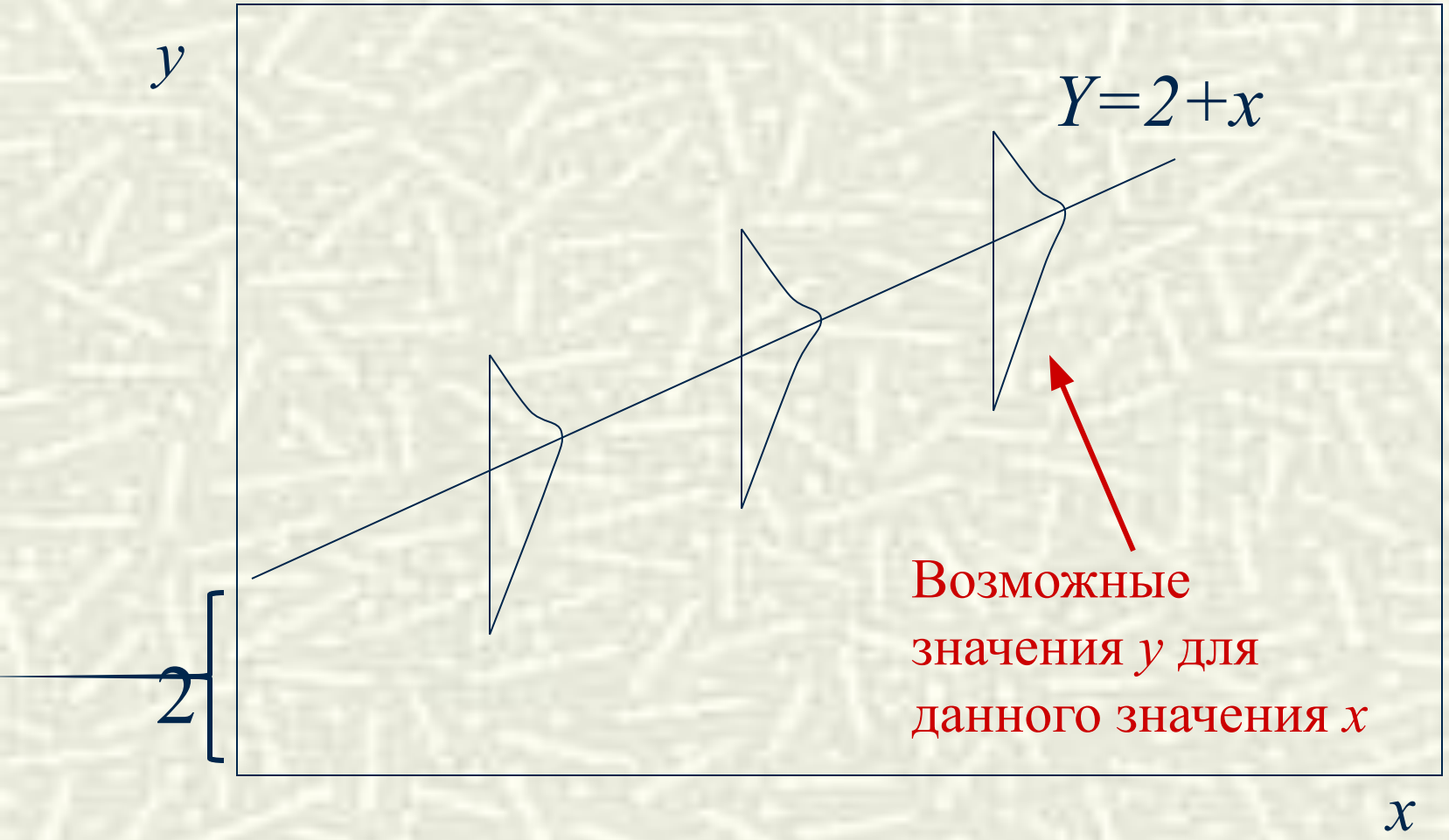
а) Корреляционные: при изменении одной из величин изменяется среднее значение другой (связь между переменными не носит направленного характера)

$M[Y/X = x] = M_x[Y] = \phi(x)$, $M[X/Y = y] = M_y[X] = \psi(y)$,

где $M[Y/X = x]$ – м. о. случайной величины Y, вычисленное при условии, что случайная величина X приняла значение x, $\phi(x) \neq \text{const}$, $\psi(y) \neq \text{const}$.

б) Регрессионные: односторонняя зависимость среднего значения случайной величины Y от одной X или нескольких $X_1,\ldots,X_m$ случайных величин.

Пример: Регрессионная зависимость



Что такое регрессионный анализ?

- Регрессионный анализ – наиболее часто используемый инструмент в эконометрике.
- Регрессионный анализ представляет собой анализ форм связи, устанавливающих количественные соотношения между случайными величинами изучаемого случайного процесса.

Определение регрессии

Регрессия – функциональная зависимость между объясняющими переменными и условным математическим ожиданием (средним значением) зависимой переменной, которая строится с целью прогнозирования этого среднего значения при фиксированных значениях объясняющих переменных.

Регрессионные модели

$M_x[Y] = \phi(X)$ – ***парная регрессия***,
$M_x[Y] = \phi(X_1,\ldots,X_m)$ – ***множественная регрессия***,
где $\phi(X) \neq const$,

X – объясняющая, входная, предсказывающая, экзогенная, неслучайная переменная, фактор, регрессор, факторный признак;

Y – зависимая, объясняемая, выходная, результирующая, эндогенная, случайная переменная, результирующий признак.

Пример: Парная регрессия

- Мы хотим определить зависимость между продажами и затратами на рекламу.
- y – продажи.
- x – рекламные расходы.

Пример: Множественная регрессия

- Мы хотим определить связь между потреблением, доходом семьи, финансовыми активами семьи и размером семьи.
- y – потребительские расходы.
- x_1 – доход семьи.
- x_2 – финансовые активы семьи.
- x_3 – размер семьи.

Регрессионные уравнения

$Y = \mathrm{M}[Y/x] + \varepsilon = \phi(x) + \varepsilon$ – уравнение парной регрессии,

$Y = \mathrm{M}[Y/x_1, \ldots, x_m] + \varepsilon = \phi(x_1, \ldots, x_m) + \varepsilon$ – уравнение множественной регрессии,

где ε – ***случайный фактор*** (***остаток***), обусловленный многими причинами.

В зависимости от вида функции $\phi(x)$ модели делятся на ***линейные*** и ***нелинейные***.

Причины обязательного присутствия случайного фактора

1. Невключение в модель всех объясняющих переменных.
2. Неправильный выбор функциональной формы модели.
3. Агрегирование переменных (факторы представляют собой комбинацию других переменных).
4. Ошибки измерений.
5. Ограниченность статистических данных.
6. Непредсказуемость человеческого фактора.

Этапы построения качественного уравнения регрессии

1. Определение конечных целей эконометрического моделирования, набора участвующих в модели факторов и их роли (***постановочный*** этап).
2. Предмодельный анализ экономической сущности изучаемого явления (***априорный*** этап).
3. Сбор необходимой статистической информации (***информационный*** этап).

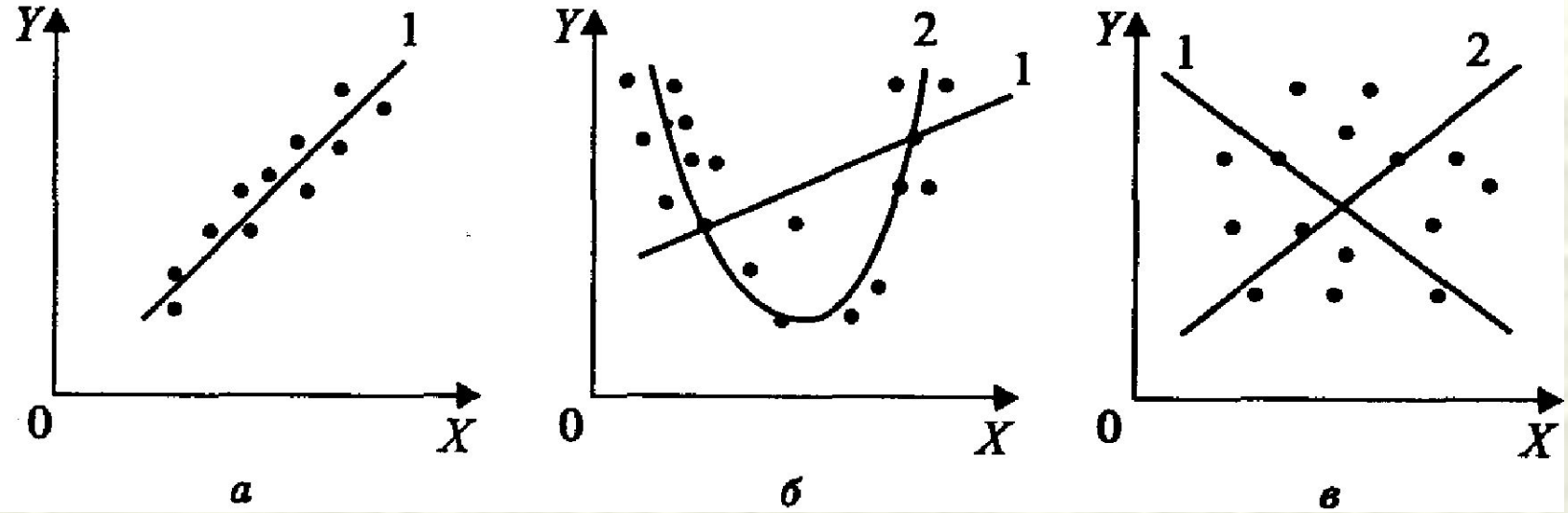
Этапы построения качественного уравнения регрессии

4. Выбор формулы уравнения регрессии (***спецификация*** уравнения регрессии).
5. Определение параметров выбранного уравнения (***параметризация***).
6. Анализ качества уравнения и проверка его адекватности эмпирическим данным, совершенствование уравнения (***верификация***).

Выбор формы парной регрессии

В случае парной регрессии выбор формулы обычно осуществляется по графическому изображению реальных статистических данных в виде точек (***корреляционное поле*** или ***диаграмма рассеивания***).

Примеры взаимосвязи между переменными



а) Взаимосвязь между Y и X близка к линейной: $Y = a + bX$

б) Взаимосвязь близка к квадратической: $Y = a + bX + cX^2$

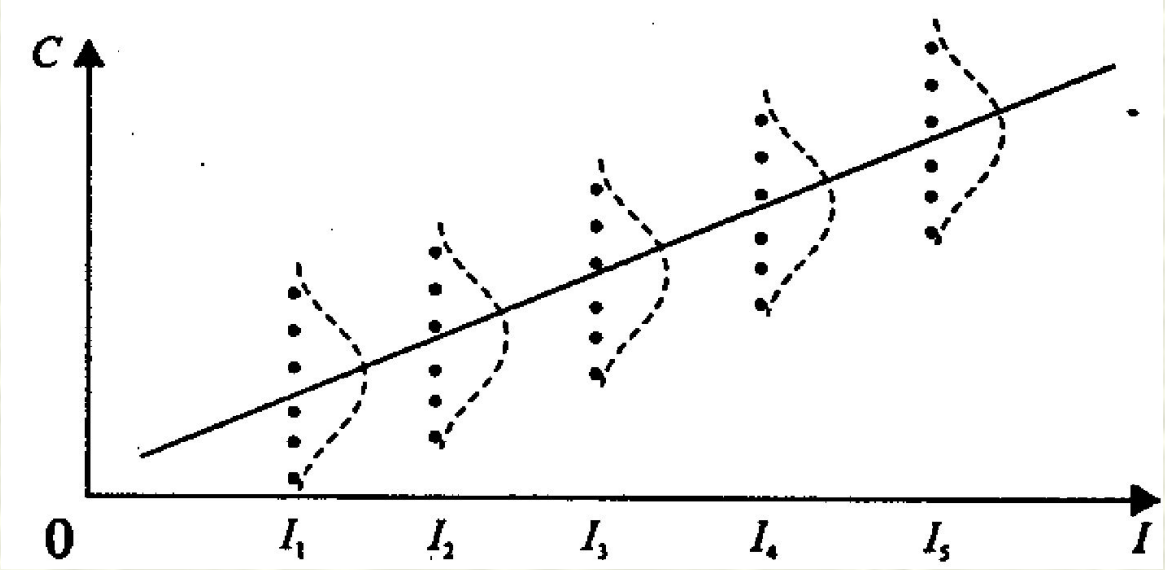
в) Взаимосвязь между Y и X отсутствует. Какую бы мы ни выбрали форму связи, результаты проверки ее качества будут неудачными

Парная линейная регрессия

Модель линейной регрессии является наиболее распространенной (и простой) зависимостью между переменными, а также может служить начальной точкой эконометрического анализа.

Модель Кейнса

Рассмотрим модель Кейнса зависимости частного потребления C от располагаемого дохода I: $C = C_0 + bI$, где C_0 – величина автономного потребления, b – предельная склонность к потреблению ($0 < b \leq 1$)



Модель парной линейной регрессии

Теоретическая парная линейная регрессионная модель:

$$y_i = M[Y / X = x_i] + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

где β_0, β_1 – теоретические коэффициенты регрессии,

ε_i – случайное отклонение.

В общем виде теоретическую парную линейную регрессионную модель будем представлять в виде:

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

Задачи линейного регрессионного анализа

Задачи линейного регрессионного анализа состоят в том, чтобы по имеющимся статистическим данным (x_i, y_i), $i = 1, 2, \ldots, n$, для переменных X и Y:

а) получить наилучшие оценки параметров β_0 и β_1;

б) проверить статистические гипотезы о параметрах модели;

в) проверить, адекватность модели данным наблюдений.

Эмпирическое уравнение регрессии

По выборке ограниченного объема нельзя точно определить теоретические значения β_0 и β_1.

Можно лишь построить ***эмпирическое уравнение регрессии***:

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

где b_0 и b_1 – оценки параметров β_0 и β_1 ***эмпирические коэффициенты регрессии***).

$\hat{y}_i$ – оценка условного м. о. $M[Y/X = x_i]$.

Эмпирическое уравнение регрессии

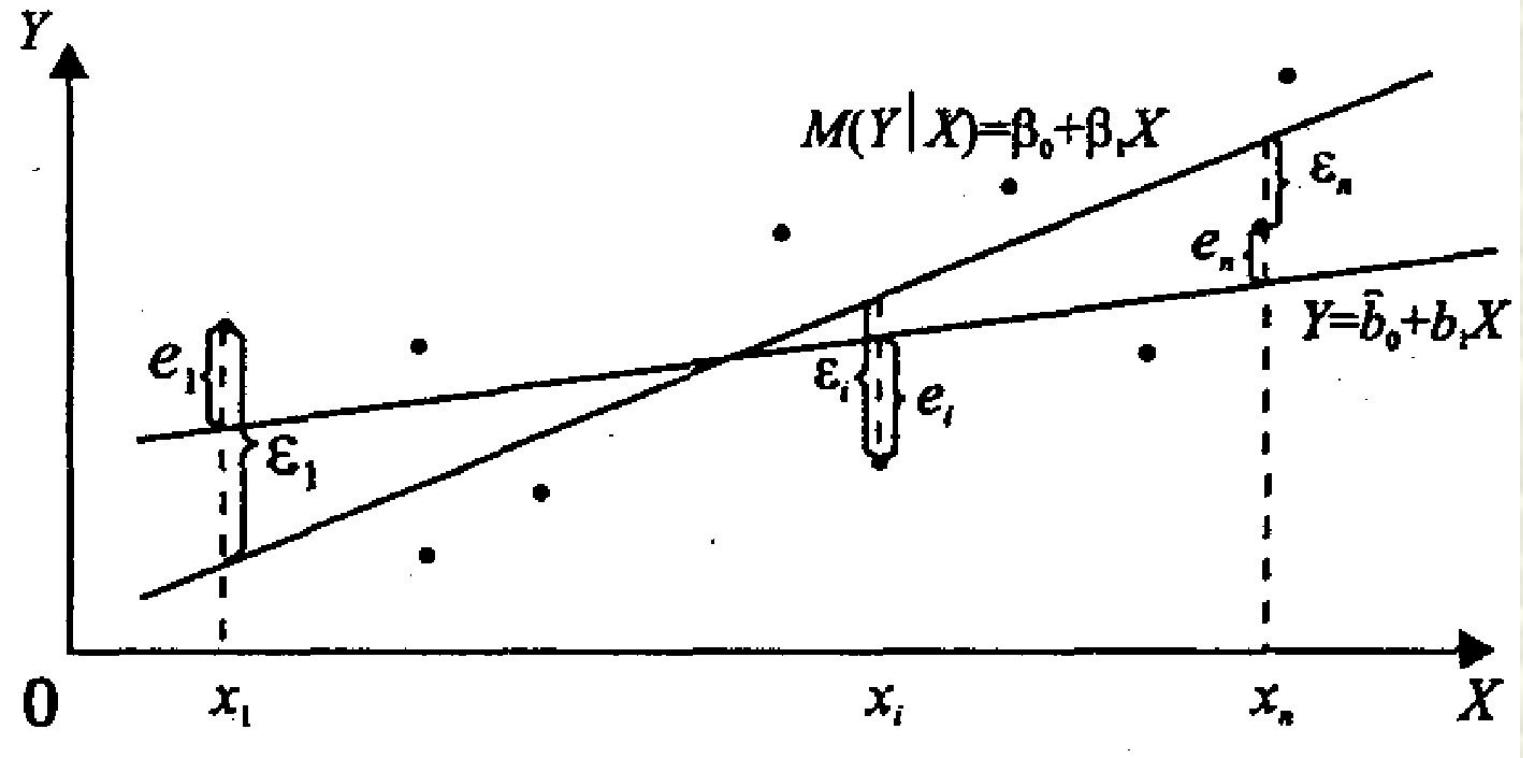
В результате имеем: $y_i = b_0 + b_1 x_i + e_i$

где e_i – оценка теоретического случайного отклонения ε_i.

Оценки b_0 и b_1 отличаются от истинных значений β_0 и β_1, что приводит к несовпадению эмпирической и теоретической линий регрессии. По различным выборкам из одной и той же генеральной совокупности получают разные значения оценок коэффициентов регрессии.

Эмпирическое и теоретическое уравнения регрессии

Соотношение между теоретическим и эмпирическим уравнениями регрессии схематично имеет вид:



Задача определения коэффициентов регрессии

Задача состоит в нахождении по выборке данных оценок b_0 и b_1 так, чтобы построенная линия регрессии была ***наилучшей в определенном смысле*** среди всех других прямых. Решение основано на минимизации:

$$g(y_i, x_i, b_0, b_1) \to \min, \quad i = \overline{1, n}$$

где g – некоторая функция.

Метод наименьших квадратов

Наиболее распространена ***методом наименьших квадратов*** (***МНК***), реализующий минимизацию суммы квадратов отклонений:

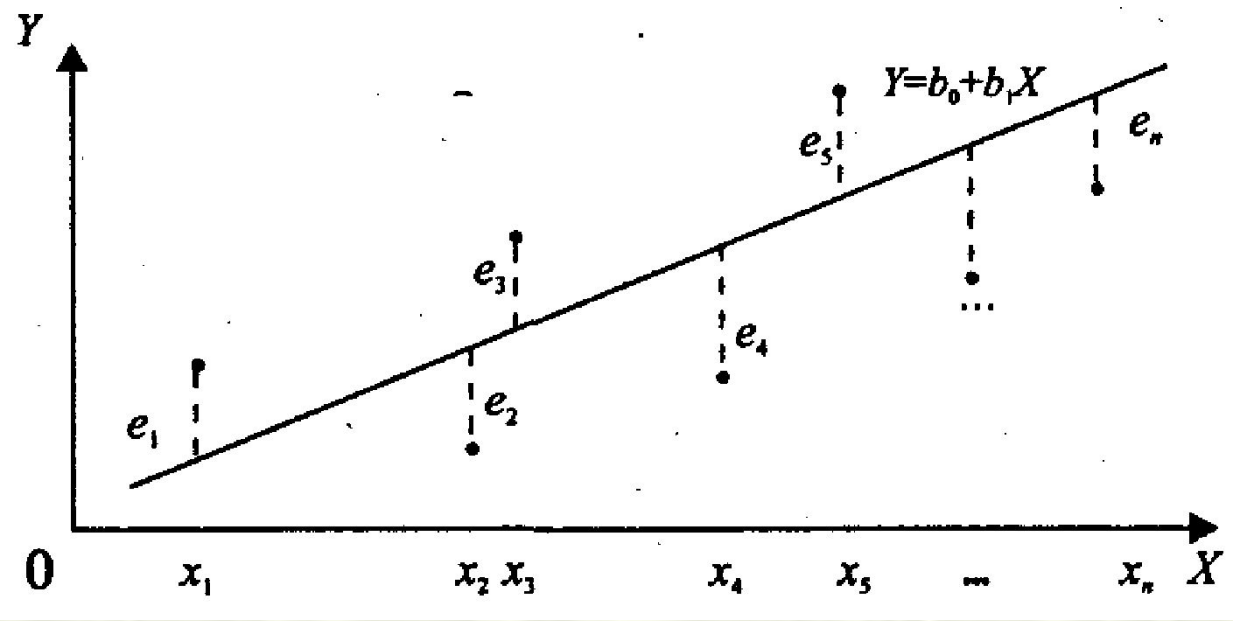
$$\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2 = \sum_{i=1}^{n} e_i^2 \to \min$$

Основные особенности МНК:

1) Он наиболее простой с вычислительной точки зрения.
2) Оценки коэффициентов регрессии по МНК при определенных предпосылках обладают рядом оптимальных свойств.

Метод наименьших квадратов

Пусть по выборке данных (x_i, y_i), $i = 1, 2, \ldots, n$, требуется определить оценки b_0 и b_1 эмпирического уравнения регрессии:



Метод наименьших квадратов

В этом случае минимизируется функция:

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

Т.к. функция $Q(b_0,b_1)$ непрерывна, выпукла и ограничена снизу, то она имеет минимум.

Необходимым условием минимума $Q(b_0,b_1)$ является равенство нулю ее частных производных по неизвестным параметрам b_0 и b_1.

Метод наименьших квадратов

Приравняем нулю частные производные и затем разделим на *n* оба уравнения:

$$\begin{cases} \dfrac{\partial Q}{\partial b_0} = -2\sum (y_i - b_0 - b_1 x_i) = 0 \\ \dfrac{\partial Q}{\partial b_1} = -2\sum (y_i - b_0 - b_1 x_i) x_i = 0 \end{cases} \Rightarrow \begin{cases} nb_0 + b_1 \sum x_i = \sum y_i \\ b_0 \sum x_i + b_1 \sum x_i^2 = \sum x_i y_i \end{cases}$$

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

Оценки метода наименьших квадратов

Решив последнюю систему уравнений, получим:

$$b_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\tilde{Cov}(x, y)}{Var(x)} = r_{xy} \sqrt{\frac{Var(y)}{Var(x)}}$$

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

Матричная форма записи

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{pmatrix} \qquad \mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} \qquad \mathbf{B} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \qquad \mathbf{e} = \mathbf{Y} - \mathbf{XB}$$

МНК эквивалентен ортогональности матрицы **X** и вектора **e**:

$$\mathbf{X}^T\mathbf{e} = \mathbf{0} \qquad \mathbf{X}^T(\mathbf{Y} - \mathbf{XB}) = \mathbf{0} \Rightarrow \mathbf{B} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Выводы

1. Оценки МНК являются функциями от выборки, что позволяет их легко рассчитать.

2. Оценки МНК являются точечными оценками теоретических коэффициентов регрессии.

3. Эмпирическая прямая регрессии обязательно проходит через точку $(\overline{x}, \overline{y})$.

Выводы

4. Эмпирическое уравнение регрессии построено так, что $\sum e_i = 0, \quad \bar{e} = 0.$

5. 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.

Другие методы определения коэффициентов регрессии

Другие методы определения коэффициентов регрессии:

- метод наименьших модулей (МНМ),
- метод моментов (ММ),
- метод максимального правдоподобия (ММП).

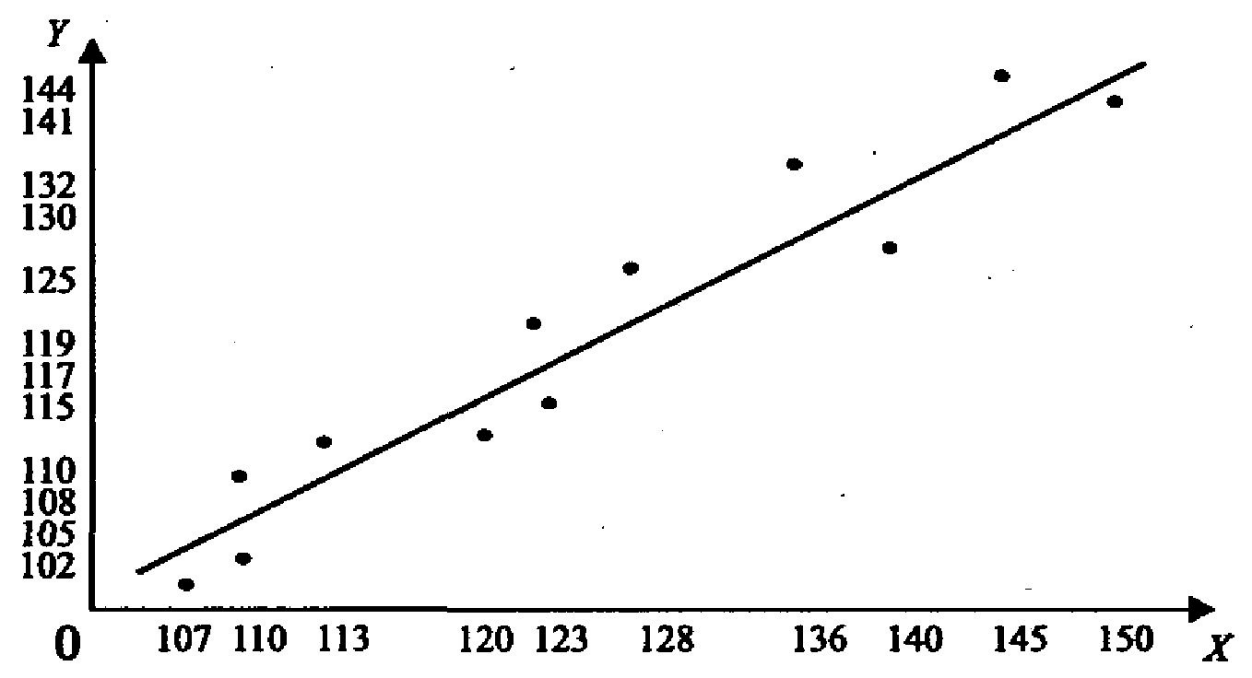
Пример (А) построения уравнения регрессии

При анализе зависимости объема потребления Y (у.е.) домохозяйства от располагаемого дохода X (у.е.) отобрана выборка объема $n = 12$ (помесячно в течение года), результаты которой приведены в таблице:

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

Пример (А) построения уравнения регрессии

Для определения вида зависимости построим корреляционное поле:



Пример (А) построения уравнения регрессии

По расположению точек на корреляционном поле делаем предположение о линейной зависимости:

$$\hat{Y} = b_0 + b_1 X.$$

Согласно МНК, имеем:

$$b_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{15298{,}08 - 125{,}25 \cdot 120{,}67}{15884{,}75 - (125{,}25)^2} = \frac{184{,}583}{197{,}188} = 0{,}9361$$

$$b_0 = \bar{y} - b_1 \bar{x} = 120{,}67 - 0{,}9361 \cdot 125{,}25 = 3{,}423$$

Пример (А) построения уравнения регрессии

Т.о., уравнение парной линейной регрессии имеет вид:

$$\hat{Y} = 3{,}423 + 0{,}9361X$$

Изобразим данную прямую регрессии на корреляционном поле. По этому уравнению рассчитаем $\hat{y}_i$, а также $e_i = y_i - \hat{y}_i$. Для анализа степени линейной зависимости вычислим:

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

Отсюда можно сделать вывод о сильной прямой линейной зависимости между переменными.

Пример (А). Таблица расчетов по МНК

№№	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	11449	10914	10404	103,58	-1,583	2,507
2	109	105	11881	11445	11025	105,46	-0,455	0,207
3	110	108	12100	11880	11664	106,39	1,609	2,587
4	113	110	12769	12430	12100	109,20	0,800	0,641
5	120	115	14400	13800	13225	115,75	-0,752	0,566
6	122	117	14884	14274	13689	117,62	-0,624	0,390
7	123	119	15129	14637	14161	118,56	0,440	0,193
8	128	125	16384	16000	15625	123,24	1,759	3,094
9	136	132	18496	17952	17424	130,73	1,270	1,614
10	140	130	19600	18200	16900	134,47	-4,474	20,015
11	145	141	21025	20445	19881	139,15	1,846	3,407
12	150	144	22500	21600	20736	143,83	0,165	0,027
Сумма	1503	1448	190617	183577	176834	-	$1{,}4 \cdot 10^{-14}$	35,249
Среднее	125,25	120,67	15884,75	15298,1	14736,2	-	-	-

Конец лекции