Тема 3. Метод наименьших квадратов

1. Спецификация линейной модели парной регрессии.
2. Оценки параметров линейной регрессии. Метод наименьших квадратов (МНК).
3. Предпосылки МНК и свойства МНК-оценок.
4. Интервалы прогноза по линейному уравнению регрессии.
5. Нелинейная парная регрессия, ее линеаризация и применение.

1 вопрос

Суть регрессионного анализа

Корреляционный анализ

- X и Y равноценны, не делятся на независимую и зависимую
- Измеряют наличие и силу взаимосвязи между X и Y, основной мерой является коэффициент корреляции

Регрессионный анализ

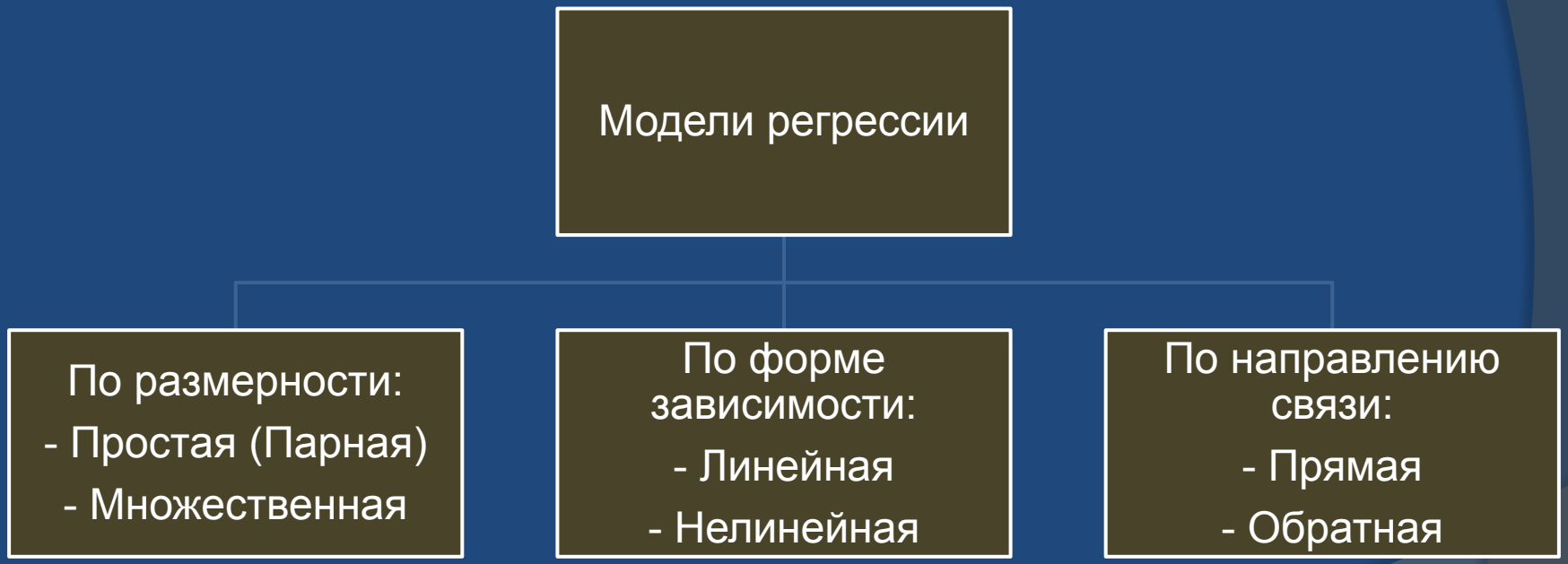
- X и Y не равноценны, изменение независимой X служит причиной для изменения зависимой Y
- Анализируют как X влияет на Y «в среднем» и определяют функцию регрессии Y на X.

Цель регрессионного анализа

Термин «регрессия» был введен Фрэнсисом Гальтоном в конце 19 века.

Основная цель – оценка функциональной зависимости между независимыми переменными X и условным математическим ожиданием (средним значением) зависимой переменной Y

Виды регрессии



Простая (парная) регрессия представляет собой модель, где среднее значение зависимой переменной Y рассматривается как функция одной независимой переменной X:

$$Y_x = f(x)$$

Множественная регрессия представляет собой модель, где среднее значение зависимой переменной Y рассматривается как функция нескольких независимых переменных X_1, X_2, …, :

$$Y_x = f(x_1, x_2, \ldots, x_m)$$

Спецификация модели - **формулирование вида модели, исходя из соответствующей теории связи между переменными. Исследование начинается с теории, устанавливающей связь между явлениями. (И. И. Елисеева)**

Определяется состав переменных и математическая функция для отражения связи между ними.

Спецификация линейной модели парной регрессии

$$Y_i = Y_{xi} + \varepsilon_i$$

Y_i - фактическое значение зависимой переменной Y

Y_{xi} - теоретическое (среднее) значение зависимой переменной Y, найденное из уравнения регрессии

ε_i - случайная величина (остаток регрессии)

Эмпирическое уравнение линейной регрессии

$$Y_{x_i} = a + b \cdot x_i$$

Yxi - теоретическое (среднее) значение зависимой переменной Y, найденное из уравнения регрессии

b - эмпирический коэффициент регрессии

a- эмпирический свободный коэффициент

В конкретном случае:

$$Y_i = a + b \cdot x_i + e_i$$

ei – оценка теоретического случайного отклонения ε

Теоретическая линейная модель парной регрессии

$$Y_i = \alpha + \beta \cdot x_i + \varepsilon_i$$

α – свободный коэффициент

β - коэффициент регрессии

εi – случайное отклонение (возмущение)

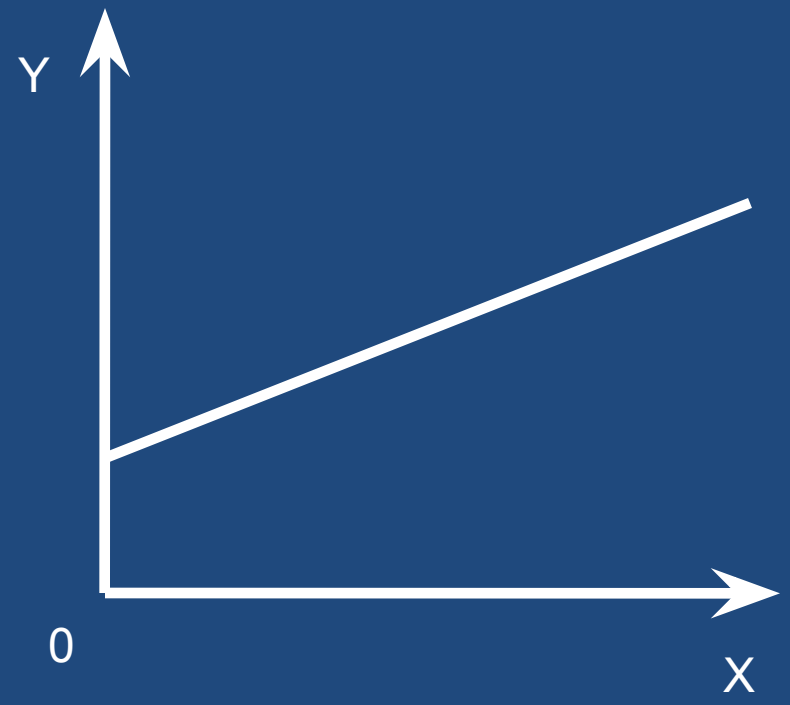
Случайное отклонение включает влияние не учтенных в модели факторов, случайных ошибок и особенностей измерения. Источники его присутствия в модели: спецификация модели, выборочный характер исходных данных, особенности измерения переменных.

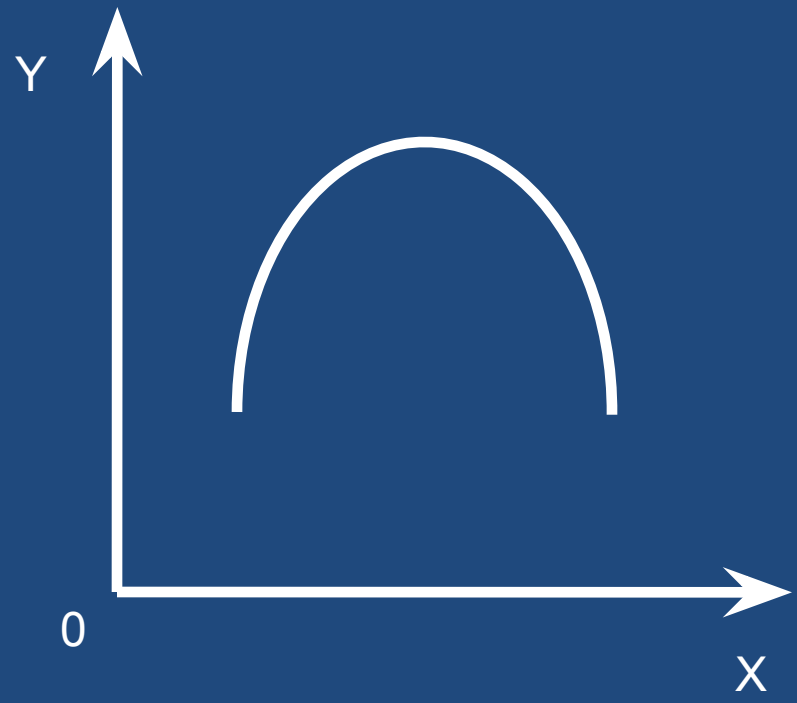
Типы ошибок в регрессии

Ошибки спецификации	Ошибки выборки	Ошибки измерения
• Неправильный выбор математической функции • Недоучет существенного фактора	• Неоднородные статистические данные • Неправильный выбор временного интервала информации	• Преднамеренные ошибки в отчетности • Непреднамеренные ошибки из-за сокрытия информации

Методы выбора типа уравнения регрессии

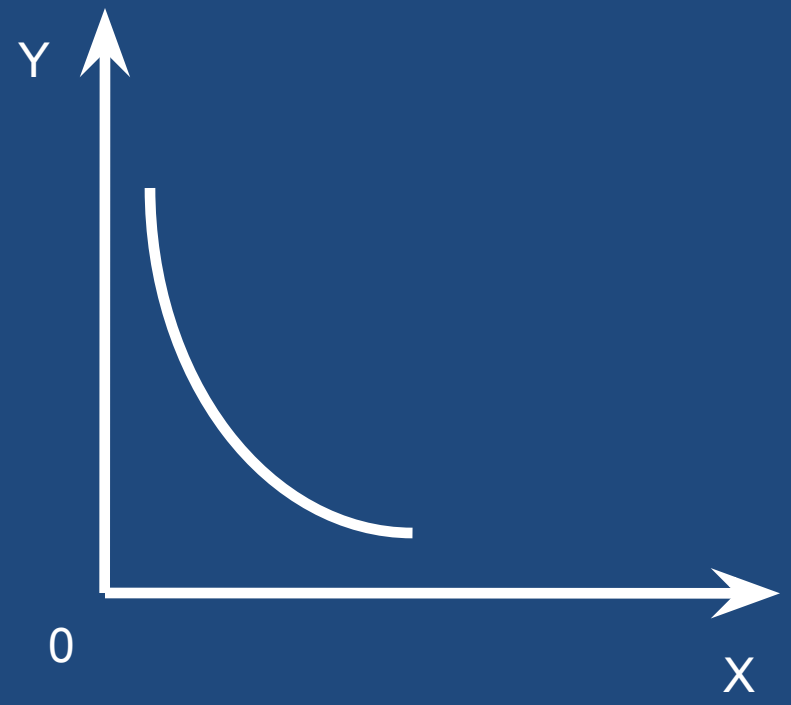
Графический метод	• Основан на визуальном анализе поля корреляции
Аналитический метод	• Основан на изучении материальной природы взаимосвязи
Экспериментальный метод	• Основан на сравнении величины остаточной дисперсии, рассчитанной при разных моделях

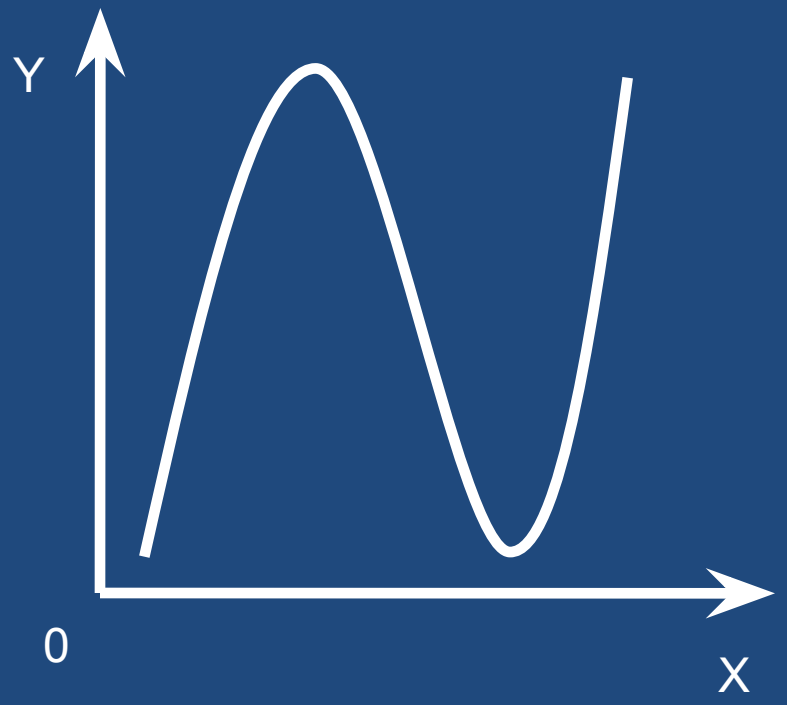




$$Y_x = a + b \cdot x$$

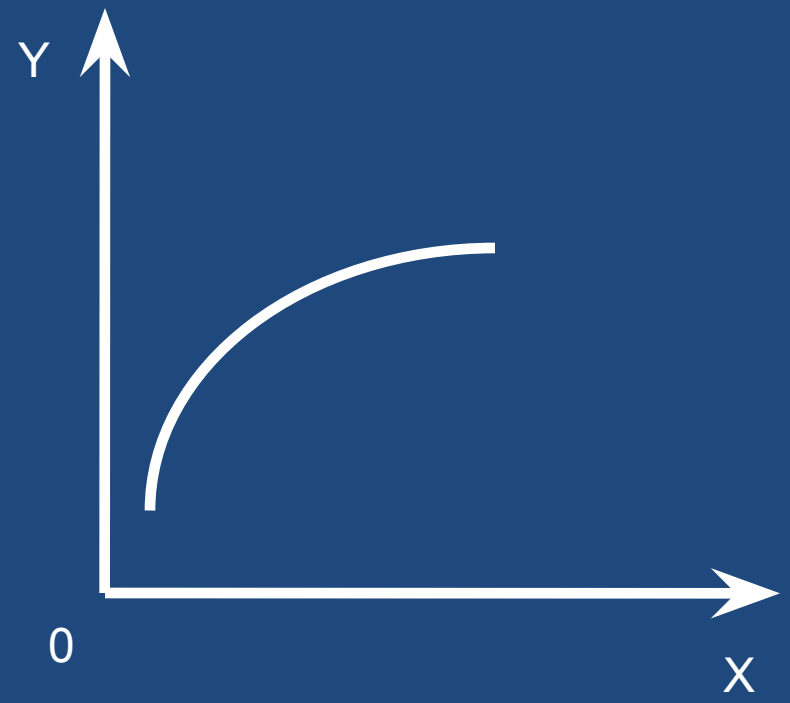
$$Y_x = a + b \cdot x + c \cdot x^2$$

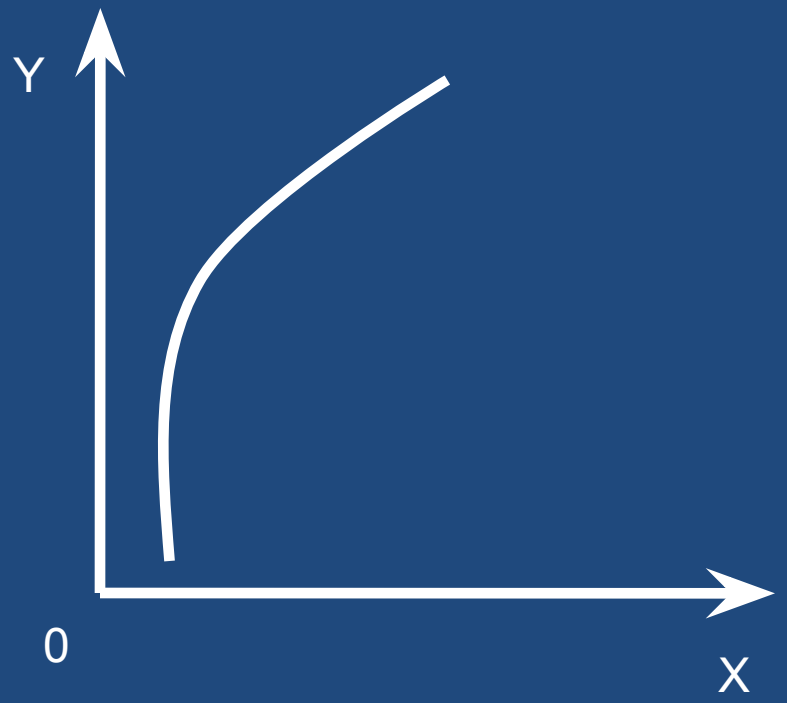




$$Y_x = a + b / x$$

$$Y_x = a + b \cdot x + c \cdot x^2 + d \cdot x^3$$





$$Y_x = a \cdot x^b$$

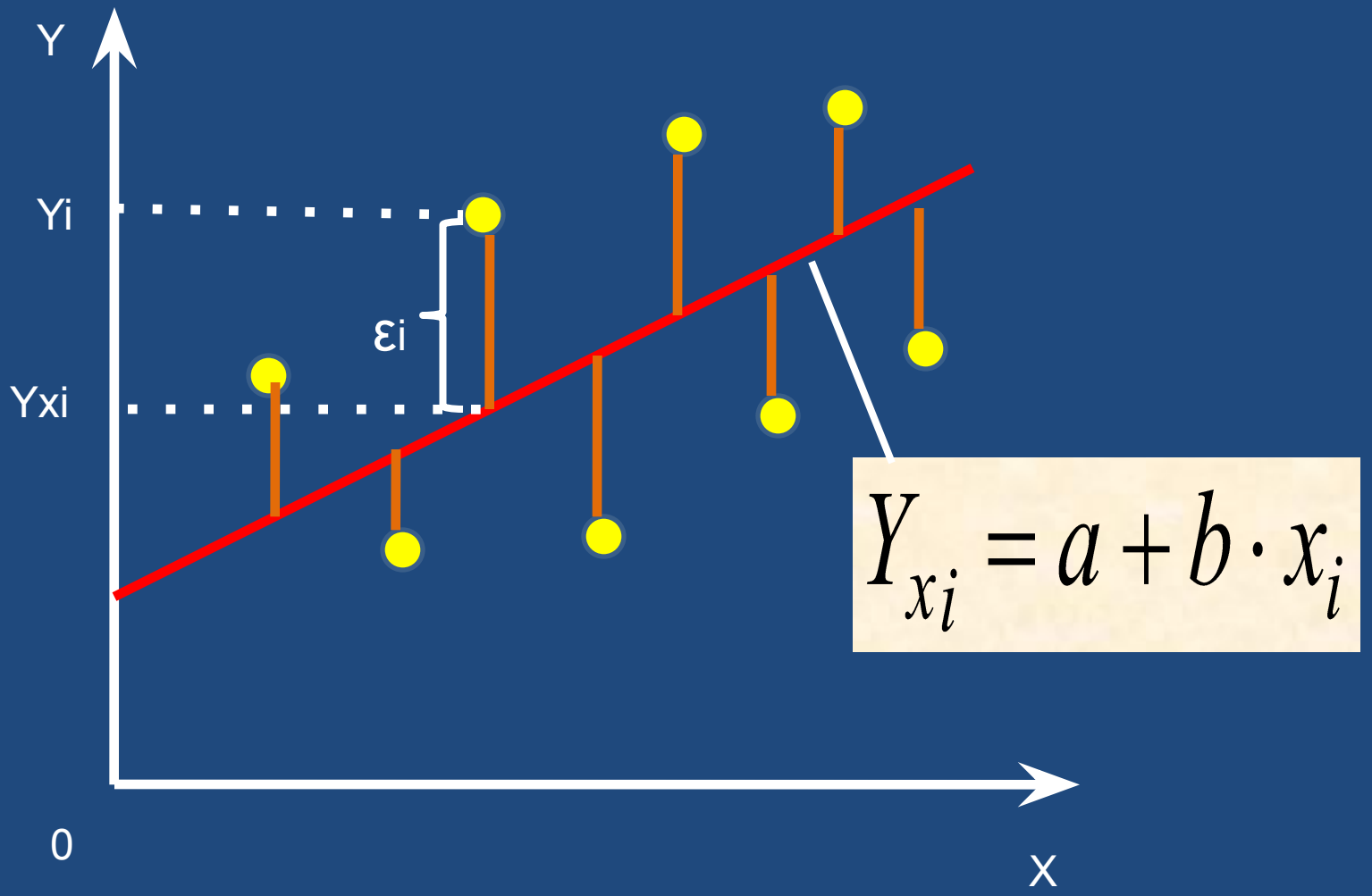
$$Y_x = a \cdot b^x$$

2 вопрос

Построение линейной регрессии сводится к оценке ее параметров – *a* и *b*

Классический подход к оцениванию параметров основан на *методе наименьших квадратов*

Из множества линий на графике выбирается та, для которой *минимальна сумма квадратов* расстояний по вертикали между точками наблюдений и этой линией



$$Y_{x_i} = a + b \cdot x_i$$

Суть метода наименьших квадратов (МНК) - оценки параметров таковы, что сумма квадратов отклонений фактических значений зависимой переменной Y от расчетных (теоретических) Y_x минимальна:

$$\sum_{i=1}^{n}(y_i - y_{xi})^2 \rightarrow \min$$

Оценка параметров регрессии

$$S = \sum (y_i - y_{x_i})^2 = \sum (y - a - b \cdot x)^2;$$

$$\frac{dS}{da} = -2\sum y + 2 \cdot n \cdot a + 2 \cdot b \sum x = 0;$$

$$\frac{dS}{db} = -2\sum y \cdot x + 2 \cdot a \sum x + 2 \cdot b \sum x^2 = 0.$$

Оценка параметров регрессии

$$\begin{cases} n \cdot a + b \sum x = \sum y, \\ a \sum x + b \sum x^2 = \sum y \cdot x \end{cases}$$

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

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