09.04.03 Прикладная информатика

Профиль «Машинное обучение и анализ данных»

Дисциплина «Математические основы анализа данных»

Лекция 9

Приближение функций
Метод наименьших квадратов

План лекции

- Постановка задачи аппроксимации
- Подходы к аппроксимации
- Интерполяция (кратко)
- Метод наименьших квадратов (МНК):
  - общие идеи метода
  - вычисление коэффициентов
  - достоинства и недостатки
  - средства Excel
  - средства Python

Постановка задачи аппроксимации

- Пусть дискретному множеству значений аргумента поставлено в соответствие множество значений функции
  $(\mathrm{i} = 0, 1, \ldots, n)$.
- Эти значения — либо результаты расчётов, либо экспериментальные данные.

x	x_0	x_1	...	x_n
$f(x)$	y_0	y_1	...	y_n

Постановка задачи аппроксимации

Аппроксимацией (от лат. proxima — ближайшая) функции называется приближённое представление сложной или заданной в виде таблицы функции **более простой** функцией $y = F(x)$, имеющей **минимальные отклонения** от исходной функции.

Подходы к аппроксимации

- **Интерполяция** – будем строить функцию $y = F(x)$, принимающую в узлах таблицы те же самые значения, т.е.

$$y_i = F(x_i) \ (i = 0,1,\ldots, n).$$

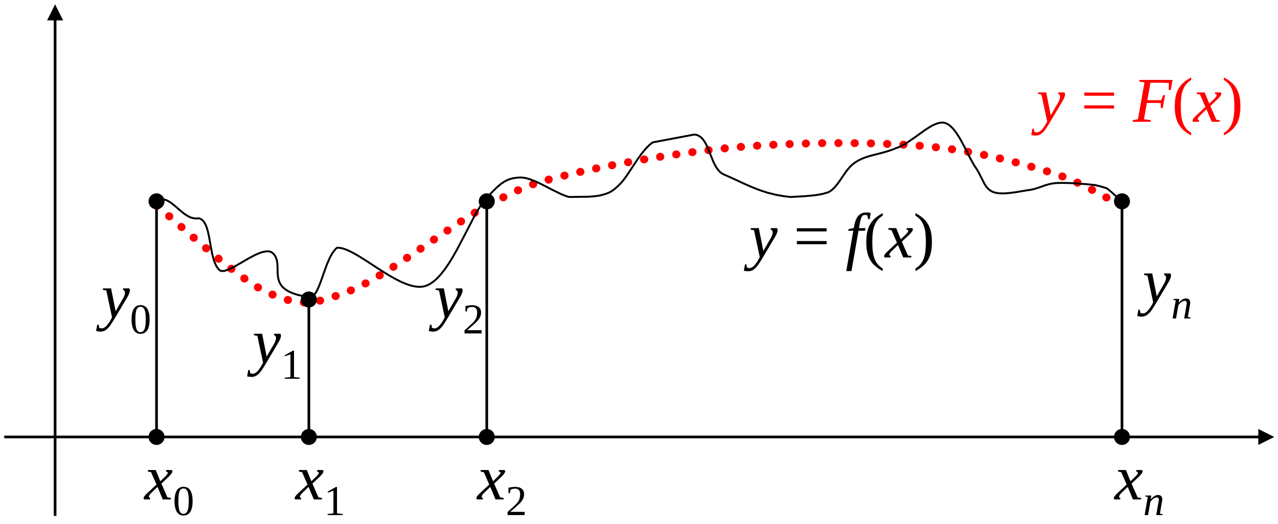
- **Методом наименьших квадратов** будем строить функцию $y = F(x)$, принимающую в узлах таблицы значения, наиболее близкие к табличным, т.е.

$$y_i \approx F(x_i)$$

Интерполяция

Требуется построить принадлежащую известному классу (имеющую простой вид) функцию $F(x)$, принимающую в точках x_0, x_1, ..., x_n те же значения, что и $f(x)$, то есть

$$F(x_0) = y_0, F(x_1) = y_1, \ldots, F(x_n) = y_n.$$



Каким образом выбрать интерполирующую функцию?

- Тригонометрические многочлены

$$P_M(x) = \frac{a_0}{2} + \sum_{k=1}^{M} (a_k \cos kx + b_k \sin kx)$$

- Алгебраические многочлены (полиномы)

$$P_n(x) = a_0 x^n + a_1 x^{n-1} + \ldots + a_n$$

- **Теорема.** Какие бы ни были заданы значения функции в $n + 1$ узлах, всегда существует и притом единственный многочлен степени не выше n, принимающий в этих узлах заданные значения.
- 100 узлов $\Rightarrow$ многочлен степени 99!

Интерполяционные многочлены

- **Лагранжа**

$$L_n(x) = \sum_{i=0}^{n} P_{n,i}(x) \cdot y_i$$

$$P_{n,i}(x) = \frac{(x - x_0)(x - x_1)\ldots(x - x_{i-1})(x - x_{i+1})\ldots(x - x_n)}{(x_i - x_0)(x_i - x_1)\ldots(x_i - x_{i-1})(x_i - x_{i+1})\ldots(x_i - x_n)}$$

- **Ньютона**

$$N_1(x) = y_0 + \frac{\Delta y_0}{1!\cdot h}\cdot(x - x_0) + \frac{\Delta^2 y_0}{2!\cdot h^2}\cdot(x - x_0)\cdot(x - x_1) +$$

$$+ \frac{\Delta^3 y_0}{3!\cdot h^3}\cdot(x - x_0)\cdot(x - x_1)\cdot(x - x_2) + \ldots + \frac{\Delta^n y_0}{n!\cdot h^n}\cdot(x - x_0)\cdot(x - x_1)\cdot\ldots\cdot(x - x_{n-1})$$

- **Гаусса**

$$P_n(x) = y_0 + q\Delta y_0 + \frac{q(q-1)}{2!}\Delta^2 y_{-1} + \frac{(q+1)q(q-1)}{3!}\Delta^3 y_{-1} + \ldots$$

$$\ldots + \frac{(q+n-1)\ldots(q-n+1)}{(2n-1)!}\Delta^{2n-1} y_{-(n-1)} +$$

$$+ \frac{(q+n-1)\ldots(q-n)}{(2n)!}\Delta^{2n} y_{-n}$$

Метод наименьших квадратов

- МНК является одним из базовых методов регрессионного анализа для оценки неизвестных параметров регрессионных моделей по выборочным данным.

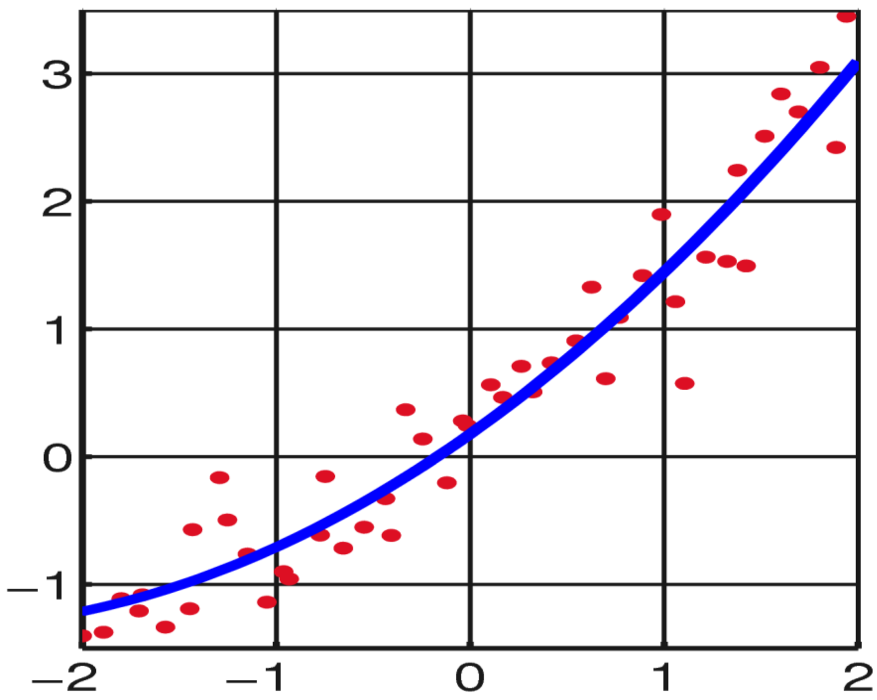
- Применяется также для приближённого представления заданной функции другими (более простыми) функциями
и оказывается полезным при обработке наблюдений.

- Пусть в результате эксперимента получена таблица значений функции y_i $(i = 1, \ldots, n)$.

x	x_0	x_1	$\ldots$	x_n
$f(x)$	y_0	y_1	$\ldots$	y_n

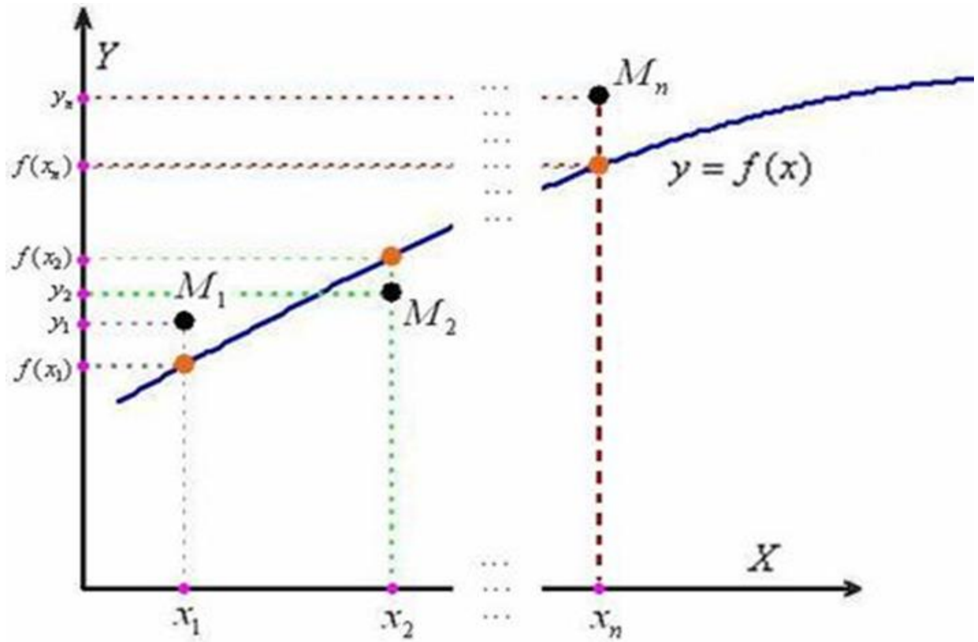
- Нам нужно подобрать функцию, **<u>график</u>** которой проходит как можно ближе к точкам таблицы.

Цель – найти формулу для описания функции, график которой проходит вблизи точек. Функций, проходящих вблизи точек, бесконечно много.



- Введём критерий близости и выберем **лучшую** функцию согласно этому критерию.

- $M_1(x_1, y_1), M_2(x_2, y_2), \ldots, M_n(x_n, y_n)$ – точки из таблицы (на графике они чёрные)
- $f(x_1), f(x_2), \ldots, f(x_n)$ – значения искомой функции в узлах x_i (оранжевые точки)



- $e_1 = y_1 - f(x_1), e_2 = y_2 - f(x_2), \ldots, e_n = y_n - f(x_n)$ – отклонения в узлах,
- $e_1^2 + e_2^2 + \cdots + e_n^2$ – сумма квадратов отклонений в узлах

Критерий близости

- Среди всех функций заданного вида наилучшей будем считать ту, которая имеет наименьшую сумму квадратов отклонений, т.е. ту, для которой

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

2 этапа нахождения аппроксимирующей функции в МНК:

1 этап – принципиально выбираем вид функции (линейная, квадратичная, степенная, логарифмическая и т.д.).

- Для этого изображаем на графике табличные значения, проводим вблизи них кривую, и в зависимости от её вида выбираем вид зависимости. Чаще всего выбирают одну из следующих приближающих функций:

1. $y = ax + b$
2. $y = ax^2 + bx + c$
3. $y = ax^m$
4. $y = ae^{mx}$
5. $y = \frac{1}{ax+b}$
6. $y = a \ln x + b$
7. $y = a\frac{1}{x} + b$
8. $y = \frac{x}{ax+b}$

Здесь a, b, c, m – параметры функциональных зависимостей.

2 этап – для функции выбранного на 1 этапе вида определяем неизвестные коэффициенты, пользуясь критерием близости

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

- Например, если на первом этапе выбрана линейная функция $y = ax + b$, то на втором этапе необходимо определить значения её коэффициентов a и b.

Вычисление коэффициентов приближающей функции

- Обозначим через *F* сумму квадратов отклонений. Легко заметить, что эта сумма зависит от коэффициентов искомой функции $F = F(a_0, a_1, a_2, \ldots)$

$$F(a_0, a_1, a_2, \ldots) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2 \to \min$$

В точке минимума функции F её частные производные обращаются в нуль (необходимое условие экстремума).

Из этих уравнений составляется система, решая которую, находят искомые коэффициенты.

$$\frac{\partial F}{\partial a_0} = 0 \qquad \frac{\partial F}{\partial a_1} = 0 \qquad \ldots \qquad \frac{\partial F}{\partial a_m} = 0$$

Поиск коэффициентов линейной функции

- Будем приближать данную табличную функцию линейной функцией вида $f(x) = ax + b$.
- В этом случае сумма квадратов отклонений будет зависеть от a и b и иметь вид $F(a, b) = \sum (y_i - (ax_i + b))^2$
- Необходимое условие экстремума: $\begin{cases} \dfrac{\partial F}{\partial a} = 0 \\ \dfrac{\partial F}{\partial b} = 0 \end{cases}$
- Составим систему относительно неизвестных коэффициентов a и b.

- Находим частные производные:

$$\frac{\partial F}{\partial a} = \left(\sum_{i=1}^{n}(y_i - (ax_i + b))^2\right)'_a = \sum_{i=1}^{n}\left[2(y_i - (ax_i + b)) \cdot (y_i - (ax_i + b))'_a\right] =$$

$$= 2\sum_{i=1}^{n}\left[(y_i - ax_i - b) \cdot (0 - (x_i + 0))\right] = 2\sum_{i=1}^{n}\left[(y_i - ax_i - b) \cdot (-x_i)\right] = 2\sum_{i=1}^{n}(ax_i^2 + bx_i - x_i y_i)$$

$$\frac{\partial F}{\partial b} = \left(\sum_{i=1}^{n}(y_i - (ax_i + b))^2\right)'_b = \sum_{i=1}^{n}\left[2(y_i - (ax_i + b)) \cdot (y_i - (ax_i + b))'_b\right] =$$

$$= 2\sum_{i=1}^{n}\left[(y_i - ax_i - b) \cdot (0 - (0 + 1))\right] = 2\sum_{i=1}^{n}(ax_i + b - y_i)$$

- Составляем систему уравнений:

$$\begin{cases} \dfrac{\partial F}{\partial a} = 0 \\ \dfrac{\partial F}{\partial b} = 0 \end{cases} \Rightarrow \begin{cases} 2\sum_{i=1}^{n}(ax_i^2 + bx_i - x_i y_i) = 0 \\ 2\sum_{i=1}^{n}(ax_i + b - y_i) = 0 \end{cases}$$

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i \end{cases}$$

- Получилась система линейных уравнений с 2мя неизвестными. Решив её любым методом, получим a и b

Алгоритм МНК для приближения линейной функцией:

1) Находим суммы $\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} y_i, \sum_{i=1}^{n} x_i^2, \sum_{i=1}^{n} x_i y_i$
2) Составляем систему уравнений с двумя неизвестными.
3) Решаем систему (например, методом Крамера)
4) Получаем искомые коэффициенты a, b, записываем функцию $f(x) = ax + b$
5) Вычисляем сумму квадратов отклонений между эмпирическими и теоретическими значениями.

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Пример:

x_i	1	2	3	4	5
y_i	5,3	6,3	4,8	3,8	3,3

- Требуется методом наименьших квадратов найти линейную функцию, которая наилучшим образом приближает эмпирические (опытные) данные.
- Коэффициенты a, b искомой приближающей функции $y = ax + b$ найдём как решение системы:

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i \end{cases}$$

Пример:

- Составим вспомогательную таблицу

x_i	1	2	3	4	5	$\sum x_i =$	15
y_i	5,3	6,3	4,8	3,8	3,3	$\sum y_i =$	23,5
x_i^2	1	4	9	16	25	$\sum x_i^2 =$	55
$x_i y_i$	5,3	12,6	14,4	15,2	16,5	$\sum x_i y_i =$	64

- Получаем следующую систему: $\begin{cases} 55a + 15b = 64 \\ 15a + 5b = 23{,}5 \end{cases}$

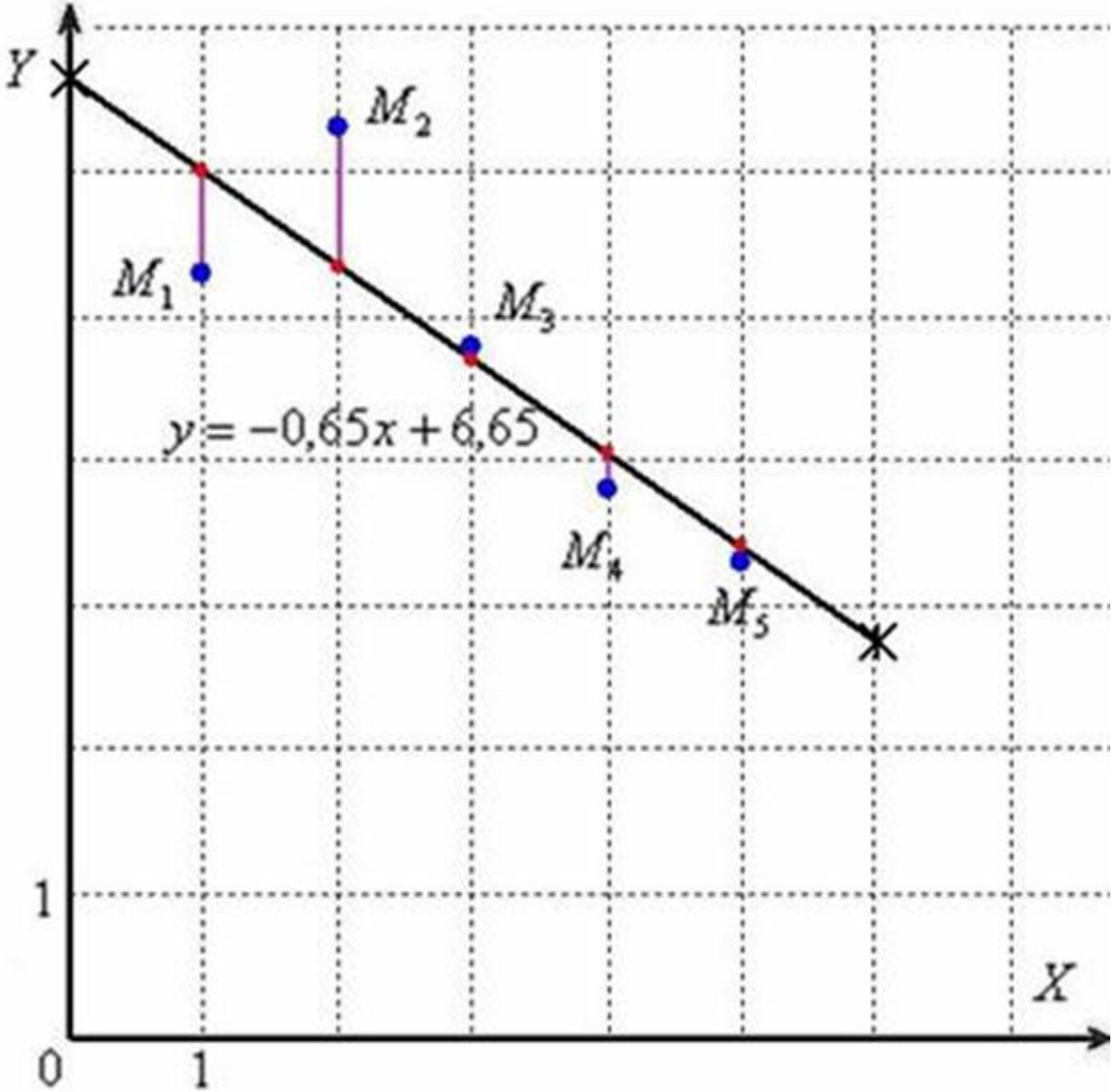
- Её решением будут значения

$$a = -0.65,\ b = 6{,}65$$

Таким образом, искомая аппроксимирующая функция:

$$y = f(x) = -0{,}65x + 6{,}65$$

- Построим график $y = -0{,}65x + 6{,}65$



- Вычислим погрешность аппроксимации

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

x_i	1	2	3	4	5		
y_i	5,3	6,3	4,8	3,8	3,3		
$f(x_i)$	6	5,35	4,7	4,05	3,4		
$(y_i - f(x_i))^2$	0,49	0,9025	0,01	0,0625	0,01	$\sum e_i^2 =$	1,475

Сведение зависимости к линейной

- Пусть требуется найти параметры зависимости вида $y = ae^{mx}$
- Сведем её к уже известной линейной:

$$\ln y = \ln ae^{mx}$$

Используем свойства логарифма:

$$\ln y = \ln a + \ln e^{mx}$$

$$\ln y = \ln a + mx \cdot \ln e$$

$$\ln y = \ln a + mx$$

$$y^* = b^* + a^* x$$

$$\sum_{i=1}^{n} y_i \to \sum_{i=1}^{n} \ln y_i \qquad a = e^{b^*}, m = a^*$$

Достоинства и недостатки МНК

+ МНК приводит к сравнительно простому математическому способу определения параметров a, b, c, … искомого функционала;

+ даёт довольно веское теоретическое обоснование с вероятностной точки зрения.

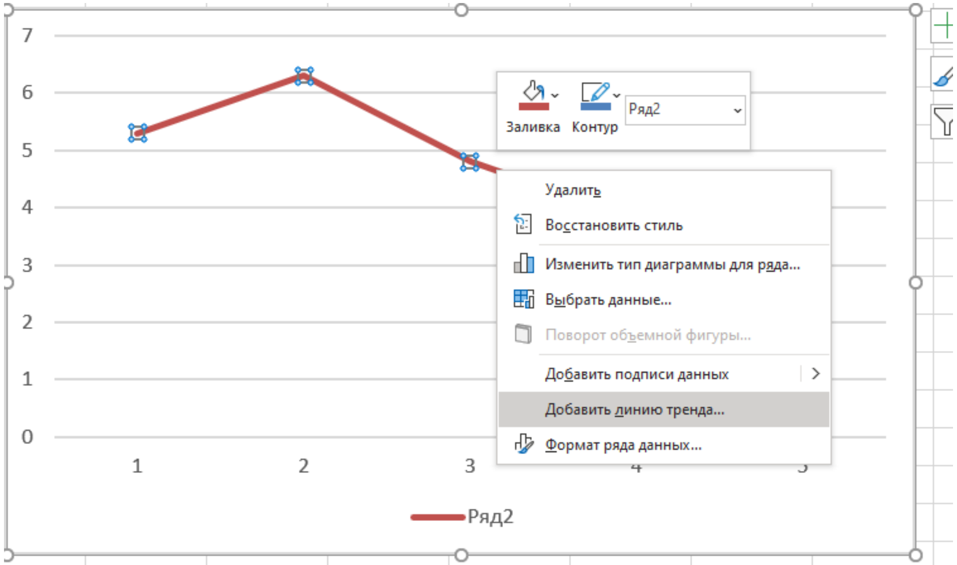
- основным недостатком МНК является чувствительность оценок к резким выбросам, которые встречаются в исходных данных.

Поиск зависимости с помощью Excel

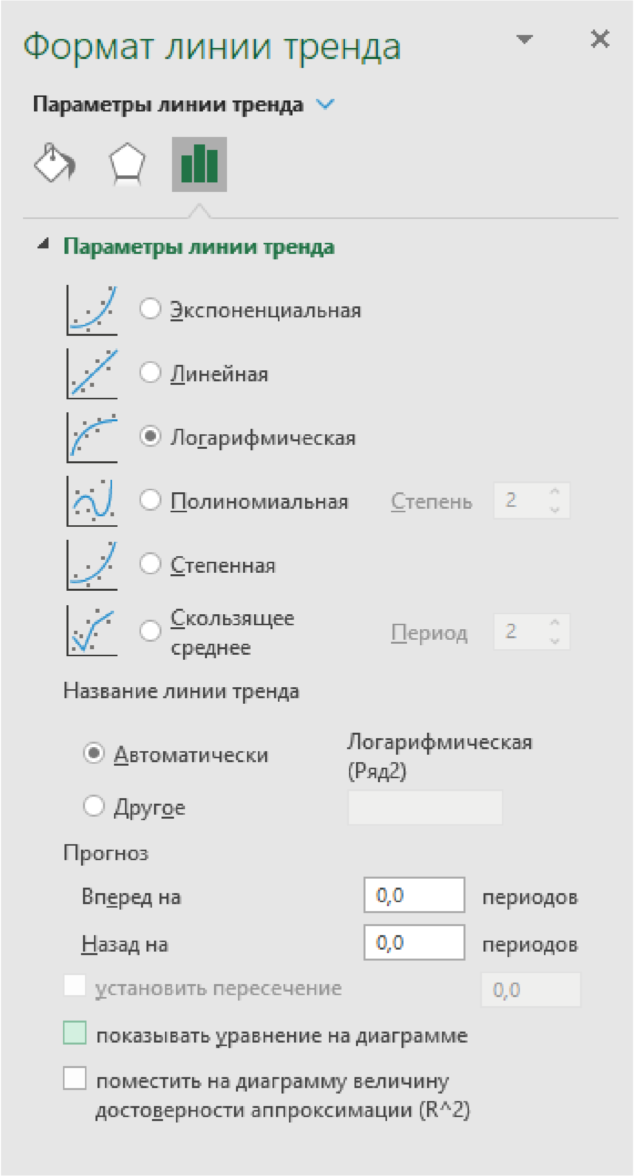
1. Строим график исходной табличной функции
2. ПКМ по графику, в контекстном меню выбираем «Добавить линию тренда»
3. В появившемся справа меню выбираем параметры зависимости, которую хотим построить. Чтобы увидеть функцию, нужно выбрать «Показывать уравнение на диаграмме». Величину погрешности показывает R^2

Поиск зависимости с помощью Excel

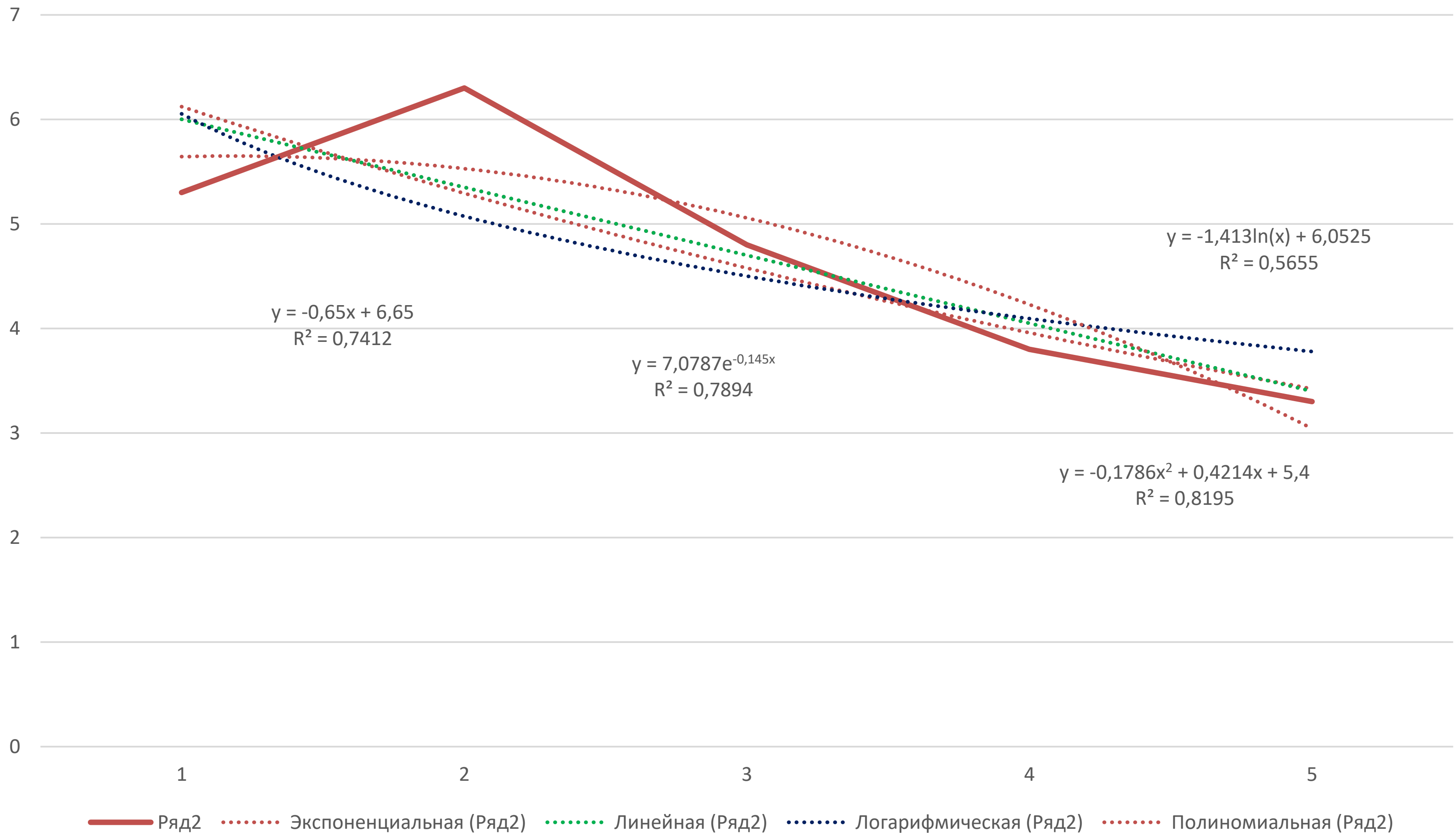
Шаг 2



Шаг 3



Результат аппроксимации в Excel



Справочная информация

Поиск коэффициентов линейной функции

- Будем приближать данную табличную функцию линейной функцией вида $f(x) = ax + b$.
- В этом случае сумма квадратов отклонений будет зависеть от a и b и иметь вид $F(a, b) = \sum (y_i - (ax_i + b))^2$
- Необходимое условие экстремума: $\begin{cases} \dfrac{\partial F}{\partial a} = 0 \\ \dfrac{\partial F}{\partial b} = 0 \end{cases}$
- Составим систему относительно неизвестных коэффициентов a и b.