Лабораторная работа №5. Обработка экспериментальных данных в электронных таблицах

Задание

1. На первом рабочем листе документа ввести исходные данные, соответствующие варианту задания. Построить график экспериментальных точек. Проанализировать экспериментальную зависимость. Выполнить подбор линейной зависимости – рассчитать коэффициенты регрессии и коэффициент корреляции. Вычислить суммарную ошибку для линейной зависимости. Построить в одной графической области экспериментальные точки и линию регрессии.
2. На втором листе вычислить коэффициенты функциональной зависимостей, соответствующей варианту задания. Расчет коэффициентов произвести при помощи метода наименьших квадратов, сведя задачу к задаче оптимизации. Для решения задачи в Excel применить инструмент Поиск решения. Построить в одной графической области экспериментальные точки и графики подобранных функциональных зависимостей. Определить суммарную ошибку и индекс регрессии.
3. На третьем рабочем листе построить линии тренда различных типов, для полиномиальной зависимости вторую и третью степени полинома. Из построенных зависимостей выбрать наилучшую.
4. Озаглавить рабочие листы согласно тематике вычислений.

ВАРИАНТЫ ЗАДАНИЙ

Вариант №1. $P(s)=As^3+Bs^2+Cs+D$

S	0,00	1,00	1,50	2,00	2,50	3,00	3,50	4,00	4,50	5,00
P	10,00	50,10	39,58	15,40	23,68	33,60	57,78	100,90	149,50	256,00

Вариант № 2. $G(s)=Ae^{bs}$

s	0,5	1,5	2	2,5	3	3,5	4	4,5	5
G	3,99	5,65	6,41	7,71	11,215	17,611	27,83	38,19	39,3

Вариант № 3. $K(s)=As^b$

s	0,5	1	1,5	2	2,5	3	3,5	4
K	1,65	2,1	2	2,1	2,3	2,4	2,22	2,59

Вариант № 4. $V(s)=As^be^{Cs}$

s	0,2	0,7	1,2	1,7	2,2	2,7	3,2
V	2,3198	2,9569	2,3999	6,4357	6,5781	6,9459	14,6621

Вариант № 5. $W(s)=1/(As+B)$

s	1	2	3	4	5	6	7	8	9
W	0,529	0,298	0,267	0,171	0,156	0,124	0,1	0,078	0,075

Вариант № 6. $Q(s)=As^2+Bs+C$

s	1	1,25	1,5	1,75	2	2,25	2,5	2,75	3
Q	5,21	4,196	3,759	3,672	4,592	4,621	5,758	7,173	9,269

Вариант № 7. $Y=x/(Ax+B)$

x	3	3,1	3,2	3,3	3,4	3,5	3,6	3,7	3,8	3,9
Y	0,61	0,6	0,592	0,58	0,585	0,583	0,582	0,57	0,572	0,571

Вариант № 8. $V=1/(A+Be^{-U})$

U	2	2,5	3	3,5	4	4,5	5	5,5	6
V	5,197	7,78	11,14	15,09	19,24	23,11	26,25	28,6	30,3

Математическая модель поставленной задачи

Рассмотрим следующую математическую задачу. В ходе эксперимента были получены данные и занесены в таблицу (табл. 1)

Таблица 1. Экспериментальные данные

x	x_1	x_2	. . . .	x_N
y	y_1	y_2	. . . .	y_N

Необходимо построить аналитическую зависимость $y=f(x)$, которая наиболее близко описывает результаты эксперимента.

Для этого построим функцию $y=f(x, a_0\ a_1, ..., a_k)$ таким образом, чтобы сумма квадратов отклонений измеренных значений y_i от расчетных $f(x_i\ ,a_0, a_1\ , ..., a_k)$ была наименьшей (см. рис. 1).

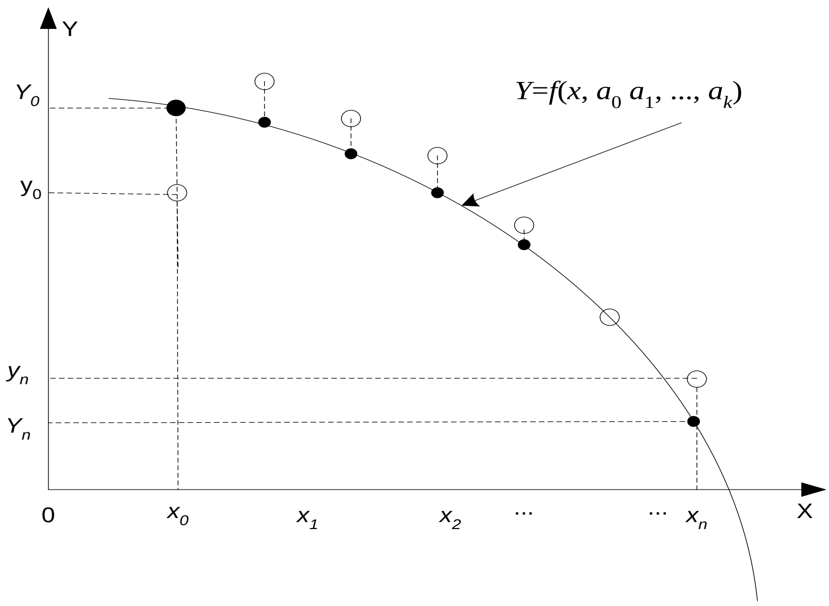


Рисунок 1. Подбор аналитической зависимости

Математически эта задача равносильна следующей – найти значение коэффициентов $a_0, a_1, a_2\ , ...,a_k$, при которых функция S (***суммарная квадратичная ошибка***) принимает минимальное значение:

$$S(a_0,a_1,..,a_k)=\sum_{i=1}^{n}[y_i-f(x_i,a_0,a_1,..,a_k)]^2\rightarrow\min \qquad (1)$$

Эта задача сводится к решению системы уравнений:

$$\begin{cases} \dfrac{\partial s}{\partial a_0} = 0; \\ \dfrac{\partial s}{\partial a_1} = 0; \\ \ldots \\ \dfrac{\partial s}{\partial a_k} = 0; \end{cases} \qquad (2)$$

Если параметры a_i входят в зависимость $y=f(x,a_o,a_1, \ ..., \ a_k)$ линейно, то мы получаем систему линейных уравнений:

$$\frac{\partial s}{\partial a_j} = 0; \quad \sum_{i=1}^{n}(-f(x_i,a_0,a_1,..,a_k))\frac{\partial f}{\partial a_j} = 0; \quad j = 0,1,..,k \qquad (3)$$

Решив систему (), найдем параметры a_o, a_1, ..., a_k и получим зависимость $y = f(x, a_o, a_1, ..., a_k)$.

Нахождение такой функции, которая была бы близка заданной, называется ***аппроксимацией*** (приближением) функции.

Линейная функция *y=ax+b* (линия регрессии). Коэффициент регрессии

Для подбора коэффициентов линейной зависимости вида $y=ax+b$ составим функцию

$$S = \sum_{i=1}^{n}\left(y_i - (ax_i + b)\right)^2 \qquad (4)$$

Коэффициенты линейной зависимости $y=ax+b$ определяются по формулам (5):

$$a = \frac{n\sum\limits_{i=1}^{n}\left(y_i x_i\right) - \sum\limits_{i=1}^{n} y_i \sum\limits_{i=1}^{n} x_i}{n\sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2}, \quad b = \frac{\sum\limits_{i=1}^{n} y_i}{n} - a\frac{\sum\limits_{i=1}^{n} x_i}{n} \qquad (5)$$

Подобранная прямая называется ***линией регрессии*** y на x, коэффициенты a и b называются ***коэффициентами регрессии***.

Чем меньше величина суммарной квадратичной ошибки $S = \sum\limits_{i=1}^{n}(y_i - (ax_i + b))^2$, тем более обосновано предположение, что табличная зависимость описывается линейной функцией.

Существует показатель, характеризующий тесноту линейной связи между x и y. Это ***коэффициент корреляции***. Он рассчитывается по формуле:

$$r = \frac{\sum\left(x_i - M_x\right)\left(y_i - M_y\right)}{\sqrt{\sum(x_i - M_x)^2\sum\left(y_i - M_y\right)^2}}, \text{где} \quad M_x = \frac{\sum x_i}{n}, \quad M_y = \frac{\sum y_i}{n} \qquad (6)$$

Значение коэффициента корреляции удовлетворяет соотношению $-1 \le r \le 1$. Чем меньше отличается абсолютная величина r от единицы, тем ближе к линии регрессии располагаются экспериментальные точки.

Если коэффициент корреляции равен нулю $r = 0$, то это только означает, что между x, y не существует линейной связи и переменные x, y называются некоррелированными. Но между этими переменными может существовать зависимость, отличная от линейной.

Если коэффициент корреляции равен нулю $r = 1$, то все экспериментальные точки ложатся на одну прямую линию.

При оценке силы связи коэффициентов корреляции используется шкала Чеддока (табл. 2).

Таблица 2. Шкала Чеддока

Значение коэффициента корреляции r	Интерпретация
$0 < r <= 0,2$	Очень слабая корреляция
$0,2 < r <= 0,5$	Слабая корреляция
$0,5 < r <= 0,7$	Средняя корреляция
$0,7 < r <= 0,9$	Сильная корреляция
$0,9 < r <= 1$	Очень сильная корреляция

Криволинейная корреляция

Если же связь криволинейная, то для оценки тесноты связи между подобранной зависимостью с экспериментальными значениями пользуются *индексом корреляции*, который рассчитывается по формуле:

$$R = \sqrt{1 - \frac{\sum_{i=1}^{n}\left(y_i - Y_i\right)^2}{\sum_{i=1}^{n}\left(y_i - M_y\right)^2}} \qquad (7)$$

где y – экспериментальные значения,
Y – теоретические значения,
M_y – среднее значение y.

Индекс корреляции по своему абсолютному значению колеблется в пределах от 0 до 1. При функциональной зависимости индекс корреляции равен 1. При отсутствии связи $R = 0$. Если коэффициент корреляции r является мерой тесноты связи только для линейной формы связи, то индекс корреляции R – и для линейной, и для криволинейной. При прямолинейной связи коэффициент корреляции по своей абсолютной величине равен индексу корреляции: $|r| = R$.

ОБРАБОТКА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ В ЭЛЕКТРОННЫХ ТАБЛИЦАХ

Подбор линейной функции $y=ax+b$ средствами MS Excel

Вычисление коэффициентов регрессии осуществляется с помощью функции:

ЛИНЕЙН(*Значения_y; Значения_x; Конст; статистика*),

где *Значения_y* — массив значений *y*,

Значения_x— необязательный массив значений *x*, если массив *x* опущен, то предполагается, что это массив {1;2;3;...} такого же размера, как и *Значения_y*.

Конст— логическое значение, которое указывает, требуется ли, чтобы константа *b* была равна 0. Если *Конст* имеет значение ИСТИНА или опущено, то *b* вычисляется обычным образом. Если аргумент *Конст* имеет значение ЛОЖЬ, то *b* полагается равным 0 и значения *a* подбираются так, чтобы выполнялось соотношение $y = ax$.

Статистика — логическое значение, которое указывает, требуется ли вернуть дополнительную статистику по регрессии. Если аргумент статистика имеет значение ИСТИНА, то функция ЛИНЕЙН возвращает дополнительную регрессионную статистику. Если аргумент статистика имеет значение ЛОЖЬ или опущен, то функция ЛИНЕЙН возвращает только коэффициенты *a* и *b*.

Коэффициенты линейной зависимости $y = ax+b$ также можно вычислить с помощью функций **НАКЛОН(***Значения_y; Значения_x***)** и **ОТРЕЗОК(***Значения_y; Значения_x***)**

Для вычисления множества точек на линии регрессии используется функция:

ТЕНДЕНЦИЯ(*Значения_y; Значения_x; Новые_значения_x; Конст*),

где *Значения_y* и *Значения_x* — массивы значений *y* и *x*, которые заданы для соотношения $y = ax + b$.

Новые_значения_x— новый массив значений, для которых ТЕНДЕНЦИЯ возвращает соответствующие значения *y*. Если этот параметр опущен, то предполагается, что новые значения совпадают с массивом значений *x*.

Конст — логическое значение, которое указывает, требуется ли, чтобы константа *b* была равна 0. Если *Конст* имеет значение ИСТИНА или опущено, то *b* вычисляется обычным образом. Если *Конст* имеет значение ЛОЖЬ, то *b* полагается равным 0, и значения *a* подбирается для соотношения $y = ax$.

Необходимо помнить, что результатом функций **ЛИНЕЙН, ТЕНДЕНЦИЯ** является множество значений – массив. То есть для ввода формулы вначале нужно выделить соответствующий по размеру диапазон ячеек, ввести формулу и нажать CTRL+SHIFT+ENTER.

Для расчета коэффициента корреляции используется функция:

КОРРЕЛ(*Массив1;Массив2*),

где *Массив1* — массив значений *x*, *Массив2* — массив значений *y*. Эти массивы должны иметь одинаковое количество точек данных.

<u>ПРИМЕР 1</u>. Зависимость $y(x)$ представлена в таблице. Построить *линию регрессии* и вычислить *коэффициент корреляции*. Найти *суммарную ошибку*. *Показать графическое решение* задачи.

y	0	0,5	1	1,5	2	2,5	3	3,5	4
x	1	2,39	2,81	3,25	3,75	4,11	4,45	4,85	5,25

1. Введем таблицу значений в лист MS Excel. Пусть исходные данные хранятся в диапазоне **B2:J3**.

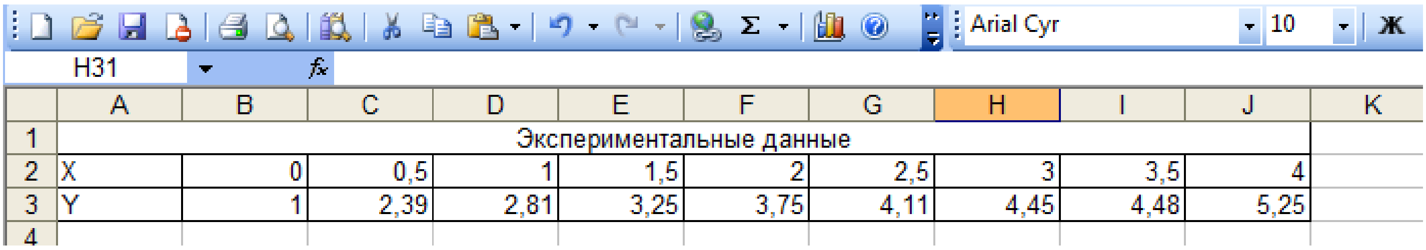


Рис. 1.

2. Выделим ячейки **A8:B8** чтобы рассчитать значения коэффициентов регрессии *a* и *b***.**

3. Обратимся к **Мастеру функций** и в категории **Статистические** выберем функцию **ЛИНЕЙН**. Заполним появившееся диалоговое окно так, как показано на рис. 3 и нажмем **OK**.

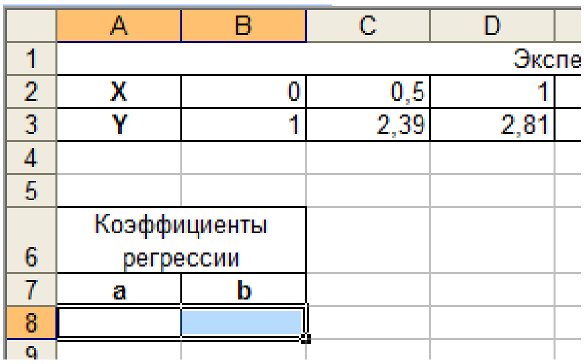


Рис. 2

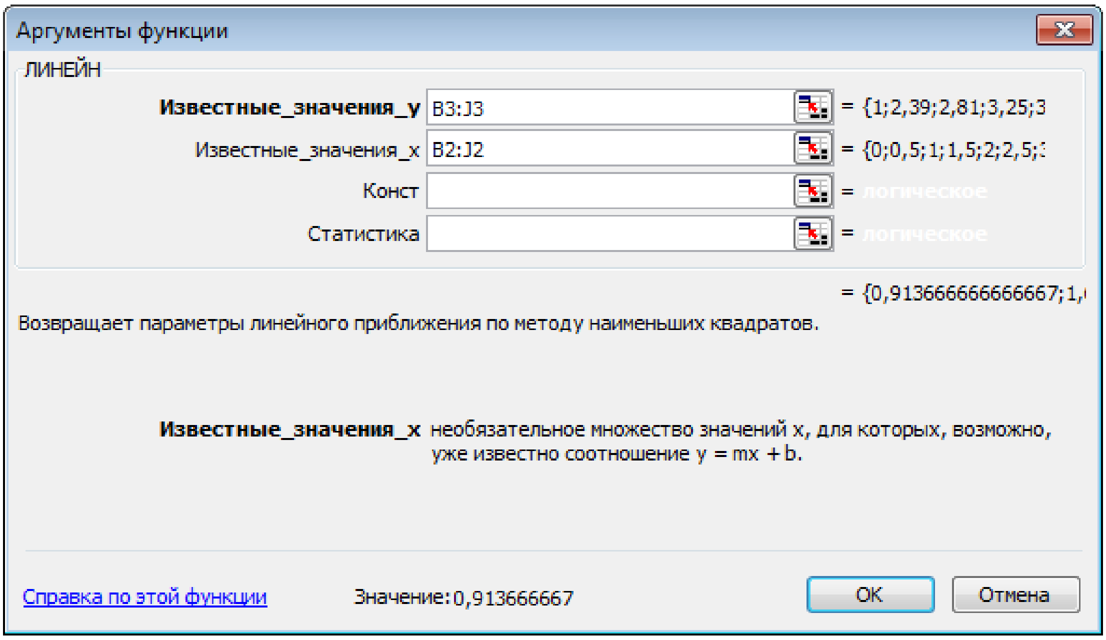
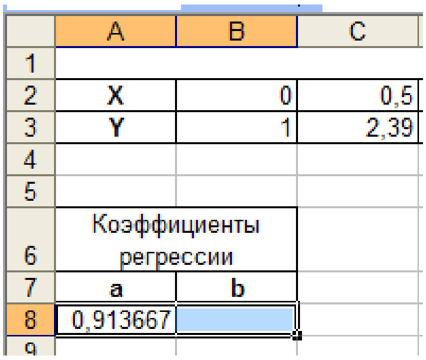


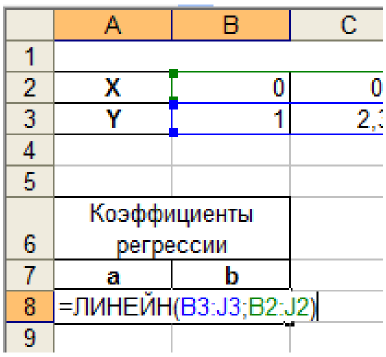
Рис. 3

4. В результате вычисленное значение появится только в ячейке **A8** (рис. 4). Для того чтобы значение появилось и в ячейке **B8** необходимо войти в режим редактирования **F2** (рис. 5), а затем нажать комбинацию клавиш **CTRL+SHIFT+ENTER** (рис. 6).



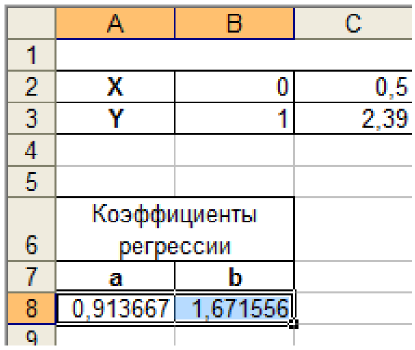
	A	B	C
1			
2	X	0	0,5
3	Y	1	2,39
4			
5			
6	Коэффициенты регрессии		
7	a	b	
8	0,913667		

Рис. 4.



	A	B	C
1			
2	X	0	0
3	Y	1	2,3
4			
5			
6	Коэффициенты регрессии		
7	a	b	
8	=ЛИНЕЙН(B3:J3;B2:J2)		
9			

Рис. 5.

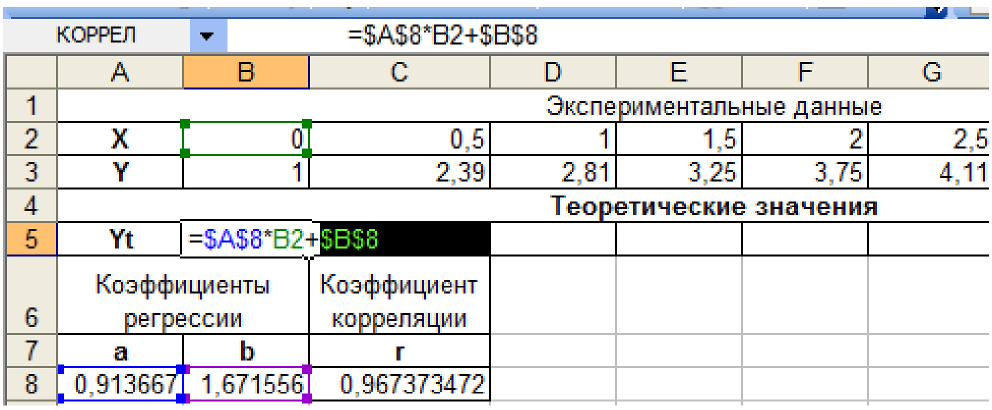


	A	B	C
1			
2	X	0	0,5
3	Y	1	2,39
4			
5			
6	Коэффициенты регрессии		
7	a	b	
8	0,913667	1,671556	

Рис. 6.

5. Зная коэффициенты регрессии *a* и *b* вычислим значения функции *y=ax+b* в заданных точках *x*. Для этого введем формулу (рис. 7)

B5=$A$8*B2+$B$8

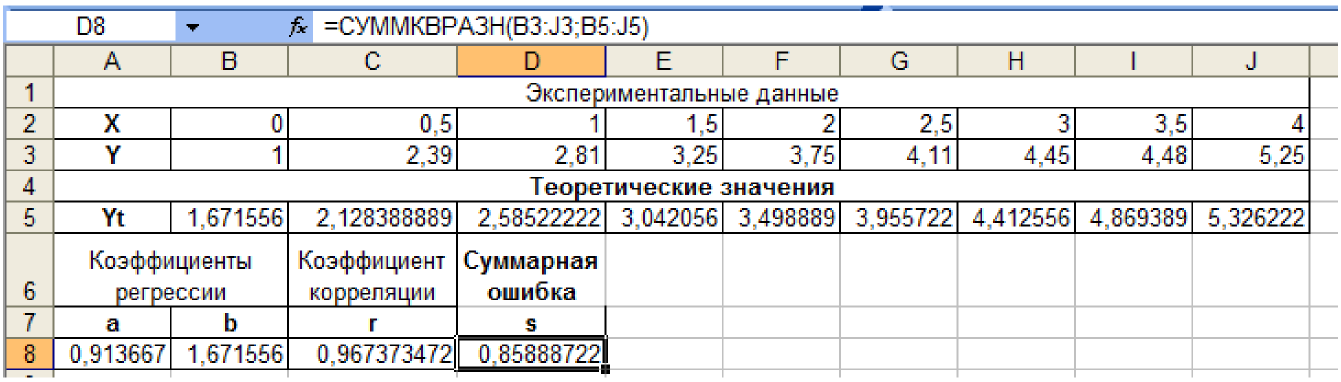


КОРРЕЛ =$A$8*B2+$B$8

	A	B	C	D	E	F	G
1				Экспериментальные данные			
2	X	0	0,5	1	1,5	2	2,5
3	Y	1	2,39	2,81	3,25	3,75	4,11
4				Теоретические значения			
5	Yt	=$A$8*B2+$B$8					
6	Коэффициенты регрессии		Коэффициент корреляции				
7	a	b	r				
8	0,913667	1,671556	0,967373472				

Рис. 7.

и скопируем ее в диапазон **C5:J5** (рис. 8).



D8 =СУММКВРАЗН(B3:J3;B5:J5)

	A	B	C	D	E	F	G	H	I	J
1				Экспериментальные данные						
2	X	0	0,5	1	1,5	2	2,5	3	3,5	4
3	Y	1	2,39	2,81	3,25	3,75	4,11	4,45	4,48	5,25
4				Теоретические значения						
5	Yt	1,671556	2,128388889	2,58522222	3,042056	3,498889	3,955722	4,412556	4,869389	5,326222
6	Коэффициенты регрессии		Коэффициент корреляции	Суммарная ошибка						
7	a	b	r	s						
8	0,913667	1,671556	0,967373472	0,85888722						

Рис. 8

6. Суммарная ошибка (1) вычисляется с помощью формулы:

D8=СУММКВРАЗН(B3:J3;B5:J5)

7. Для расчета значения коэффициента корреляции (2) была введена формула: **C8=КОРРЕЛ(B3:J3;B2:J2)**.

8. Для графической интерпретации задачи выделим диапазон экспериментальных данных **B2:J3** и, удерживая клавишу **Ctrl**, добавим диапазон теоретических значений **B5:J5**.

	A	B	C	D	E	F	G	H	I	J
1	Экспериментальные данные									
2	X	0	0,5	1	1,5	2	2,5	3	3,5	4
3	Y	1	2,39	2,81	3,25	3,75	4,11	4,45	4,48	5,25
4	Теоретические значения									
5	Yt	1,671556	2,128388889	2,58522222	3,042056	3,498889	3,955722	4,412556	4,869389	5,326222

Рис. 9.

Далее обратимся к **Мастеру диаграмм** и построим точечный график (рис. 10).

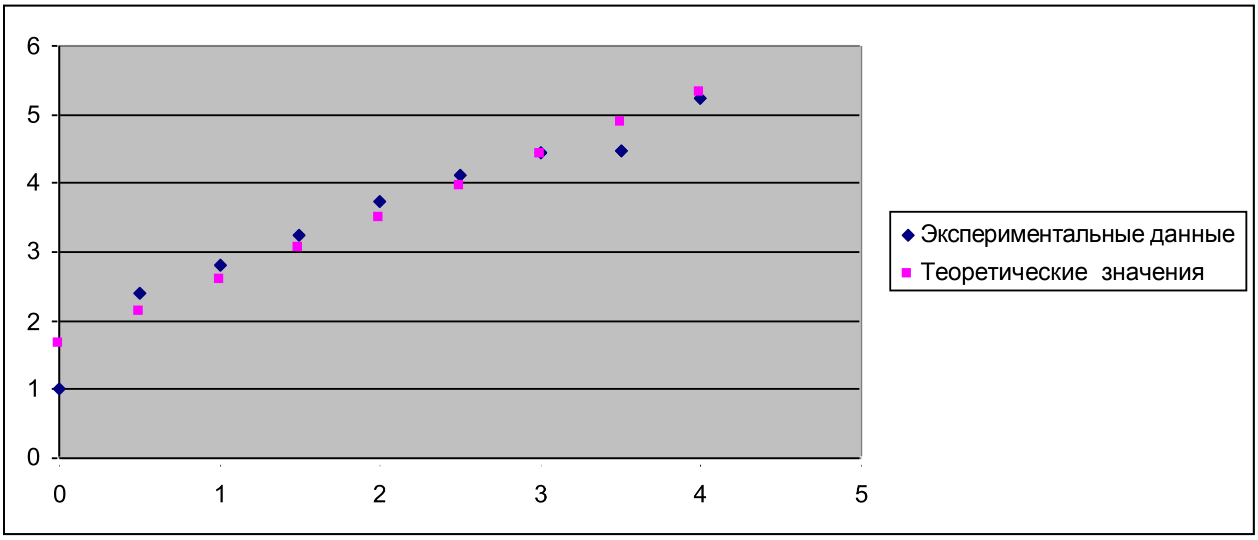


Рис. 10.

Выделим теоретические значения на графике и с помощью контекстного меню (рис. 11) изменим Тип диаграммы (рис. 12). Графическое решение задачи показано на рис. 13. Здесь изображены экспериментальные данные и подобранная к ним линия регрессии.

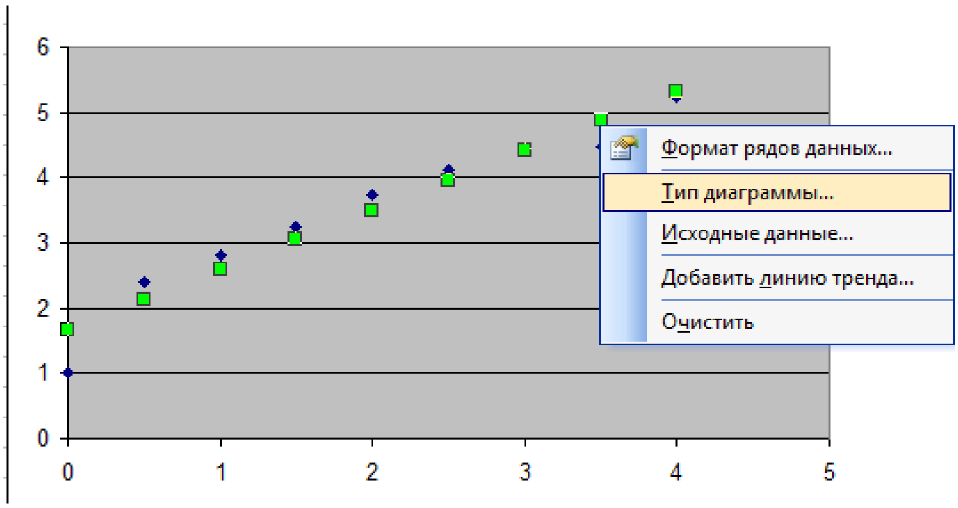


Рис. 11

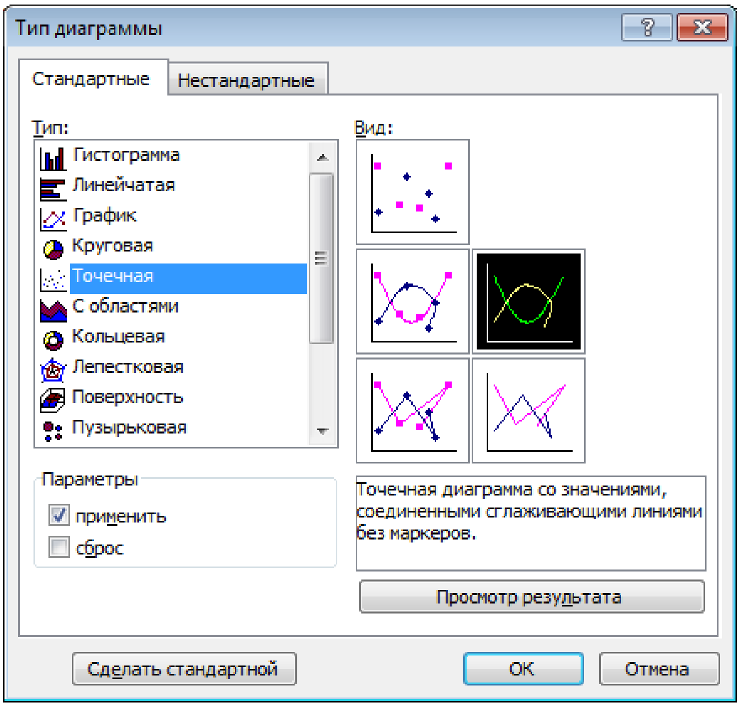


Рис. 12

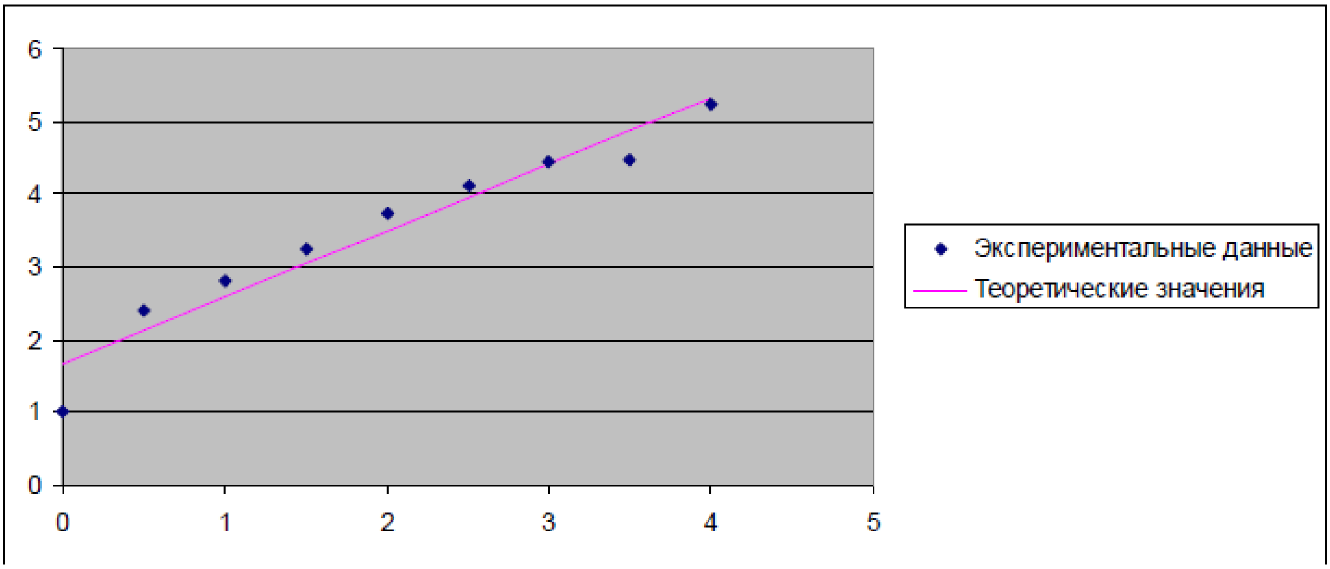


Рис. 13

Подбор нелинейных функций средствами MS Excel

Применение Поиска решения для подбора зависимости

Большинство задач, которые решается с помощью электронных таблиц, предполагают нахождение результата по известным исходным данным. Но в Excel есть инструменты, которые позволяют решить обратную задачу – подобрать исходные данные для получения желаемого результата. Для решения задач, когда требуется найти несколько параметров или комбинацию параметров, определяющих максимальное или минимальное значение в заданной ячейке, можно использовать инструмент ***Поиск решения***, который применяется для решения так называемых "задач оптимизации".

Изначально ***Поиска решения*** нет в Excel, поэтому для его использования необходимо установить соответствующую надстройку. То есть, если на вкладке ***Данные*** нет кнопки ***Поиск решения*** (рис. 14), то этот инструмент надо подключить

– откройте настройки Excel и на вкладке ***Надстройки*** в строке ***Управление*** щелкните по кнопке ***Перейти*** (рис. 15), в открывшемся диалоговом окне включите флажок ***Поиск решения***.

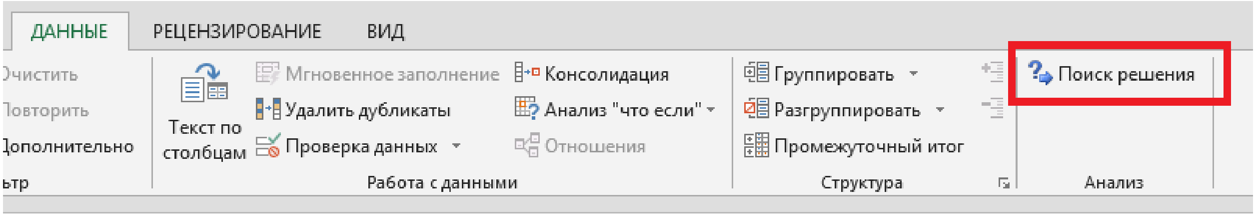


Рисунок 14. Вызов Поиска решения

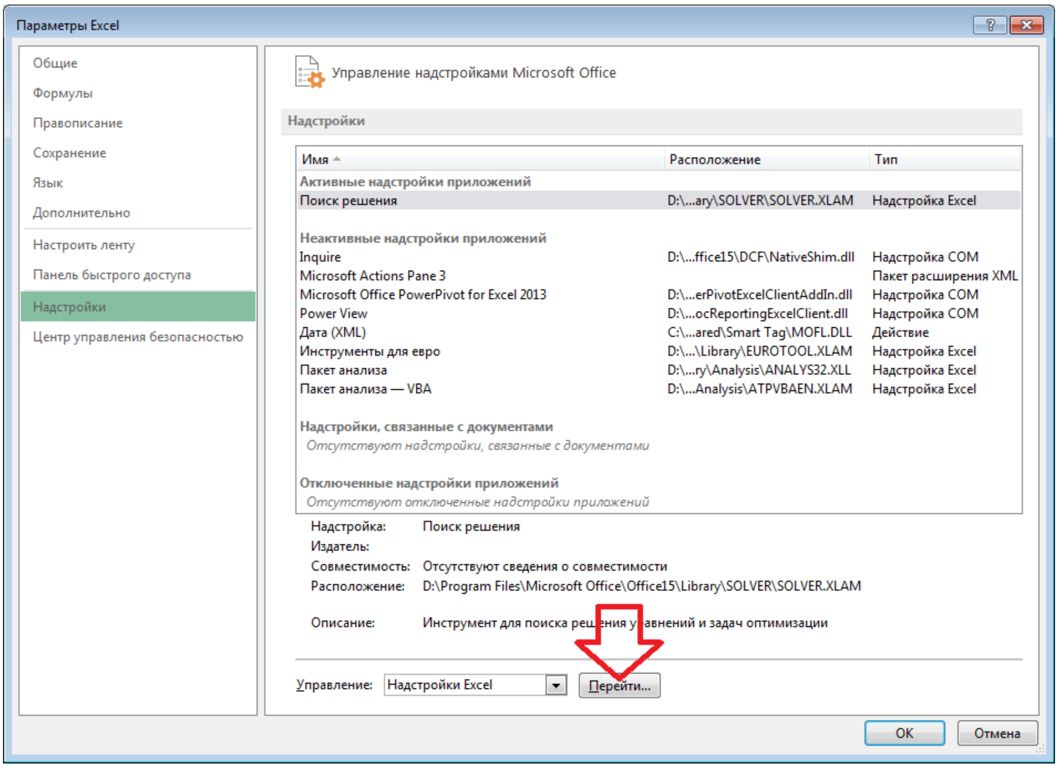


Рисунок 15. Подключение инструмента Поиск решения

<u>Пример 2</u>. Известна табличная зависимость *P(I)*. Необходимо подобрать коэффициенты зависимости вида $P(I)=AI^4+BI^3+CI+D$ с помощью метода наименьших квадратов.

I	0	0,5	1	1,5	2	2,5	3	3,5	4
P	1,1	2,35	2,81	3,25	3,75	4,11	4,45	4,84	5,25

Эта задача эквивалентна задаче нахождения минимума функции четырех переменных

$$S(A,B,C,D)=\sum_{i=1}^{9}[P_i - At_i^4 - Bt_i^3 - Ct_i - D]^2 \rightarrow \min \qquad (8)$$

Введем табличную зависимость на рабочий лист MS Excel и построим график функции (см. рис. 17)

Рассмотрим процесс решения задачи оптимизации. Пусть значения коэффициентов искомой зависимости **A**, **B**, **C** и **D** хранятся в ячейках **A6:D6**. В ячейку **B8** введем значение функции AI^4+BI^3+CI+D в первой точке (ее значение хранится в ячейке **B1**):

B8 =$A$6*B1^4+$B$6*B1^3+$C$6*B1+$D$6

Получим ожидаемое значение в точке **B1** (вначале в ячейке будет значение 0). Продублируем эту формулу на весь диапазон **B8:J8**.

В ячейку **B10** введем формулу, вычисляющую квадрат разности между экспериментальными и расчетными точками[1] :

B10 = СУММКВРАЗН(B2:J2;B8:J8)

Теперь осталось с помощью решающего блока решить задачу оптимизации без ограничений. Выделите ячейку **B10,** перейдите на вкладку Ленты ***Данные*** и вызовите инструмент ***Поиск Решения*** (рис. 15), заполните соответствующим образом появившееся диалоговое окно (рис. 16).

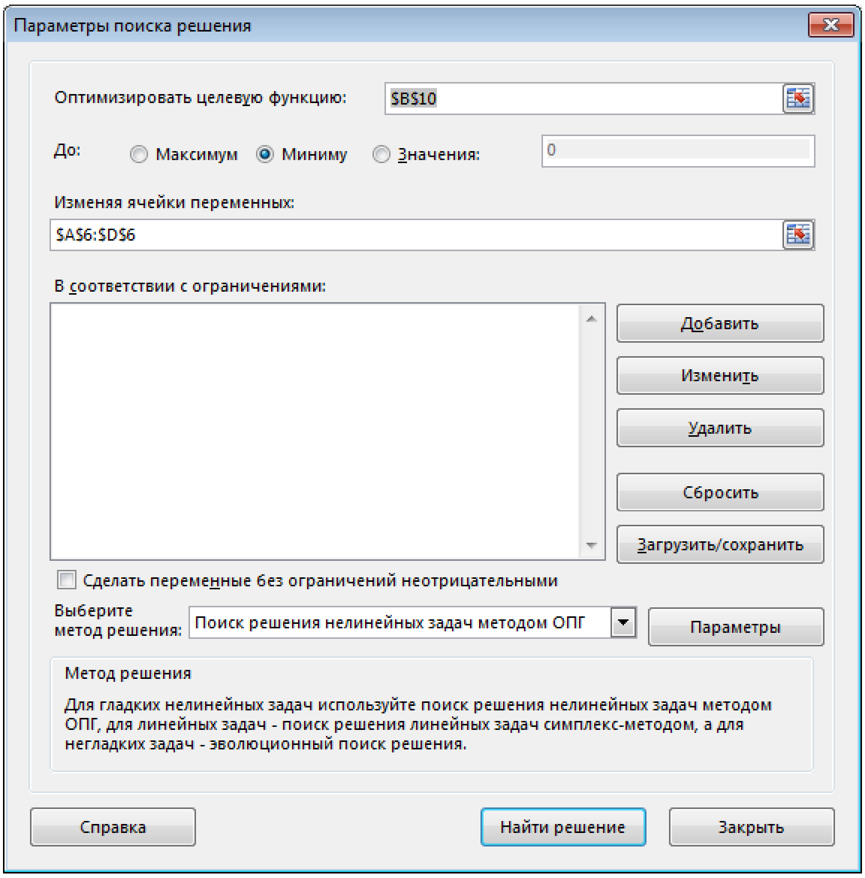


Рисунок 16. Диалоговое окно Поиска решения

В результате после нажатия кнопки **Найти решение** в ячейках **A6:D6** появятся значения коэффициентов функции At^4+Bt^3+Ct+D. В ячейках **B8:J8** получим ожидаемые значение функции в исходных точках.

Для расчета индекса корреляции (7 в ячейку **B12** введем формулу

=(1-B10/КВАДРОТКЛ(B2:J2))^0,5

В этой формуле функция **КВАДРОТКЛ** вычисляет сумму квадратов отклонений значений данного диапазона от среднего значения.

В завершении изобразим экспериментальные точки, подобранную линию и ряд прогнозируемых значений в одной графической области. Рис. отображает решение поставленной задачи. На рисунке показан лист Excel с формулами для расчета.

[1] функция **СУММКВРАЗН** находится в категории ***Математические*** мастера функций

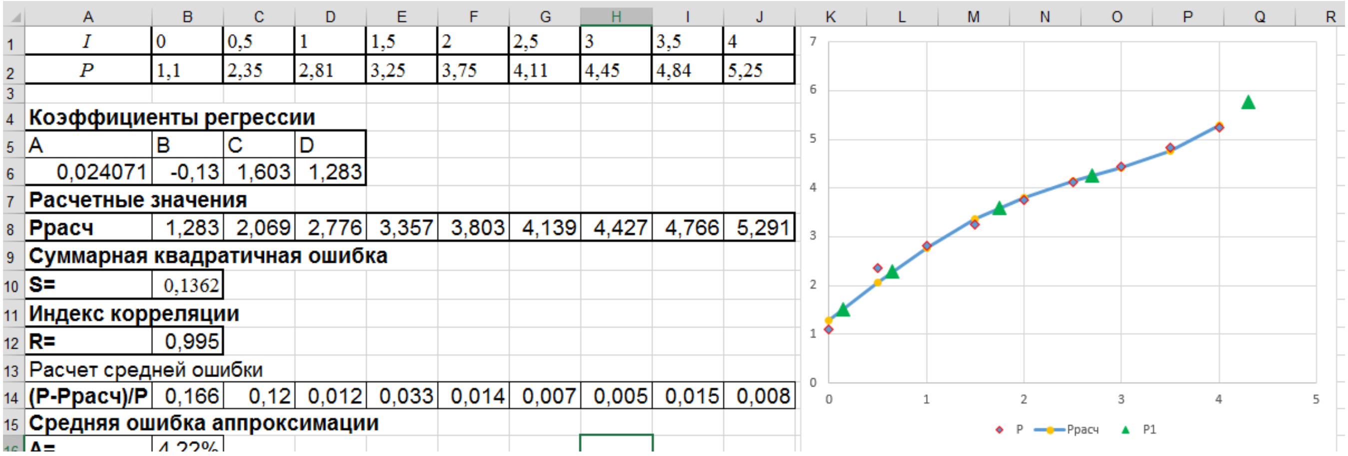


Рисунок 17. Результаты работы

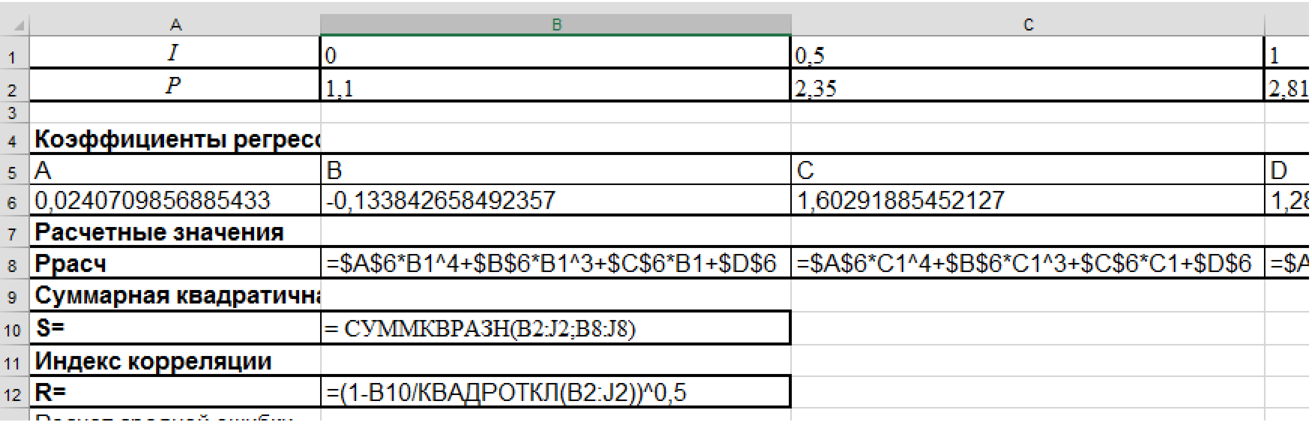


Рисунок 18. Фрагмент листа с формулами

Построение линии тренда

Построение различных аппроксимирующих зависимостей в MS Excel реализовано в виде инструмента, который размещен в свойствах диаграммы – линия тренда.

Пример 3. В результате эксперимента была определена некоторая табличная зависимость.

X	0,15	0,16	0,17	0,18	0,19	0,2
Y	4,48	4,49	5,47	6,05	6,69	7,39

Необходимо построить аппроксимирующую зависимость – полином 2-й степени. Построить графики табличной и подобранной аналитической зависимости. Вычислить ожидаемое значение в указанных точках $x_1=0,1539$, $x_2=0,2569$, $x_3=0,28$.

Введем исходные данные на лист Excel и построим точечный график (рис. 19).

X	0,15	0,16	0,17	0,18	0,19	0,2
Y	4,48	4,49	5,47	6,05	6,69	7,39

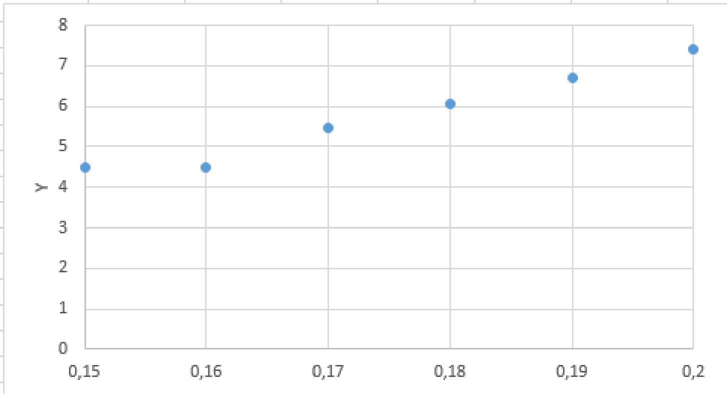


Рисунок 19. Точечный график по исходным данным

Для добавления линии тренда выделим экспериментальные точки на графике, щелкнем правой кнопкой мыши и вызовем команду **Добавить линию тренда**. Появившееся диалоговое окно (рис.) позволяет построить аппроксимирующую зависимость. Вначале нужно выбрать вид аппроксимирующей зависимости (например, выбрать полиномиальную зависимость второй степени). Далее определяются параметры построения:

- Название аппроксимирующей зависимости
- Прогноз вперед (назад) на *n* единиц – этот параметр определяет, на какое количество единиц вперед (назад) необходимо продлить линию тренда.
- Показывать ли точку пересечения кривой с прямой *Y=const;*
- Показывать аппроксимирующую функцию на диаграмме или нет;
- Помещать ли на диаграмму величину среднеквадратичного отклонения или нет (параметр поместит на диаграмму величину достоверности аппроксимации).

Результат добавления линии тренда показан на рис. 21. При построении в диалоговом окне (рис. 20) был установлен флажки *Показывать аппроксимирующую функцию* и *Помещать ли на диаграмму величину среднеквадратичного отклонения.*

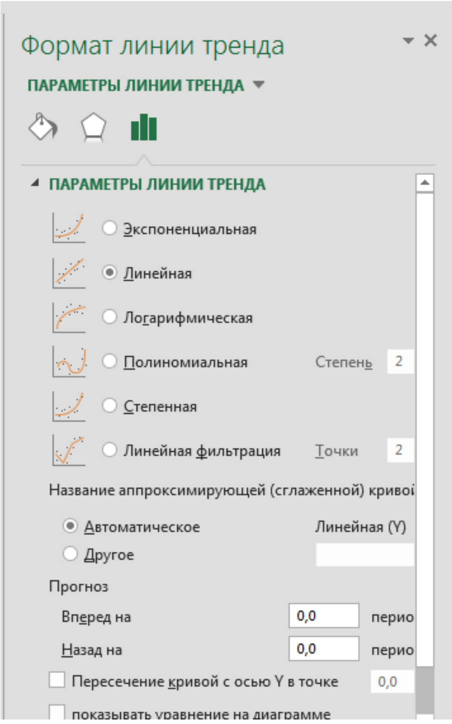


Рисунок 20. Диалоговое окно добавления линии тренда

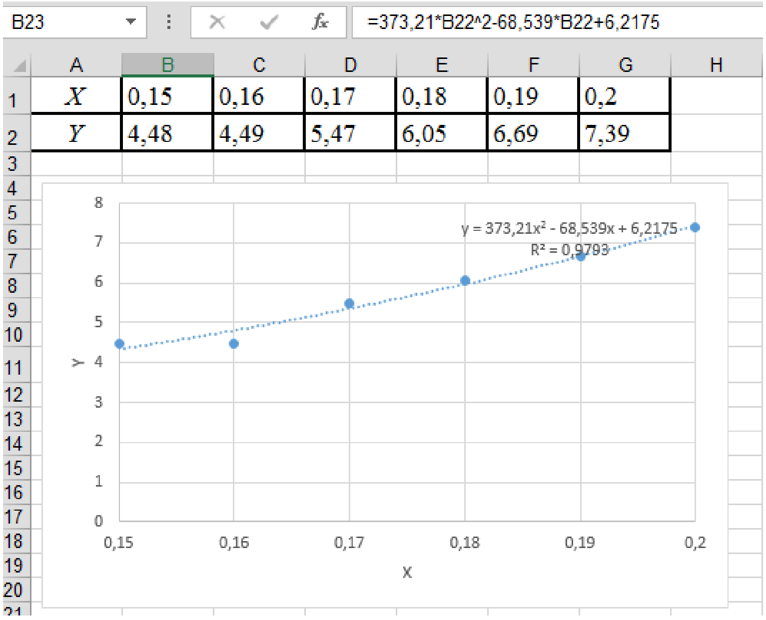


Рисунок 21. Фрагмент листа Excel с расчетами и линией тренда

Аналогично с помощью линии тренда можно подобрать и параметры других типов зависимостей (линейной, логарифмической и экспоненциальной и т.д.). По коэффициенту – величине достоверности аппроксимации можно выбрать из нескольких зависимостей наилучшую.