Лабораторная работа № 3

Расчеты коэффициента корреляции и уравнения регрессии в MS Excel

Цель работы. При помощи возможностей *Microsoft Excel* научиться обрабатывать статистические данные и находить между ними зависимость.

Краткие теоретические сведения. Для выполнения заданий необходимо воспользоваться уравнением среднеквадратической регрессии:

$Y = m_y + r\frac{\sigma_y}{\sigma_x}(X - m_x)$, (**1**)

где m_x- математическое ожидание X;

m_y- математическое ожидание Y;

r- коэффициент корреляции величин X и Y.

$r = \frac{\mu_{xy}}{\sigma_x \sigma_y}$ (**2**)

μ_{xy}- корреляционный момент величин X и Y;

$\sigma_x \sigma_y$- дисперсии X и Y.

Порядок выполнения работы.

Для приведенных в таблице результатов эксперимента определить параметры линейной регрессии, т.е. коэффициенты уравнения $y = ax + b$. Построить на одной координатной плоскости два графика зависимости y от x: экспериментальный и полученный в результате линейного приближения, т.е. прямую $y = ax + b$.

1. Согласно варианту задания построить два столбца экспериментальных данных. Варианты заданий показаны в таблице 1:

Таблица 1

i	Значения $y_i=y(x_i)$									
0	1	2	3	4	5	6	7	8	9	10
1	5,998	6,030	5,85	6,310	5,650	6,323	3,88	4,08	3,90	4,03
2	5,820	6,072	5,619	6,308	5,431	6,523	3,86	4,18	3,83	4,23
3	5,754	6,297	5,569	6,546	5,250	6,646	3,84	4,38	3,60	4,49
4	5,828	6,428	5,426	6,855	5,000	7,256	3,91	4,46	3,47	4,71
5	5,627	6,425	5,237	7,073	4,790	7,487	3,71	4,44	3,31	5,00
6	5,597	6,473	5,025	7,770	4,569	7,827	3,49	4,55	3,05	5,26
7	5,693	6,592	4,988	7,225	4,296	8,133	3,51	4,66	3,14	5,36
8	5,469	6,815	5,037	7,739	4,065	8,402	3,68	4,89	2,83	5,87

i	Значения $y_i=y(x_i)$									
0	1	2	3	4	5	6	7	8	9	10
9	5,413	6,786	4,586	7,995	3,837	8,581	3,74	4,86	2,66	5,67
10	5,526	6,925	4,575	8,063	3,519	9,014	3,47	5,04	2,53	5,89
11	5,344	7,116	4,445	8,247	3,281	9,049	3,60	5,22	2,35	6,16
12	5,304	7,053	4,353	8,472	2,926	9,571	3,51	4,99	2,49	6,65
13	5,352	7,224	3,933	8,627	2,801	9,891	3,48	5,39	2,19	6,39
14	5,301	7,439	3,899	8,936	2,546	10,073	3,30	5,56	1,82	6,81
15	5,424	7,302	3,793	9,082	2,232	10,406	3,23	5,42	1,69	7,08
16	4,996	7,426	3,473	9,076	2,016	10,821	3,26	5,85	1,54	7,24
17	5,080	7,97	3,551	9,363	1,794	11,151	3,14	5,99	1,22	7,61
18	5,256	7,871	3,171	9,679	1,663	11,232	3,17	5,85	1,17	7,64
19	5,090	7,929	3,330	9,846	1,375	11,655	2,96	6,01	1,04	8,03
20	5,053	8,060	3,044	10,013	1,217	11,952	2,81	5,97	1,12	10,013

2 При помощи функции КОРРЕЛ (), которая находится в разделе Статистические, определить коэффициент корреляции между величинами X и Y. Сделать вывод о зависимости случайных величин.

3 При помощи функции ЛИНЕЙН () определить параметры линейного приближения по методу наименьших квадратов. Для этого необходимо выделить две ячейки, в которых будут находиться значения параметров a и b, вызвать функцию ЛИНЕЙН (), подставить значения X, Y, константа - 1, Стат - 0. Нажать одновременно клавиши Shift, Ctrl, Enter для получения коэффициентов уравнения $y= ax +b$.

Таблица 2

Результаты корреляции

Коэффициент корреляции	Параметр *a*	Параметр *b*
–0,7144	–0,576	5,994

4. Построить столбец теоретических значений Y, подставив полученные параметры в уравнение линейной зависимости $y=ax+b$. Результаты экспериментальных и теоретических значений показаны в таблице 3.

Таблица 3

Результаты экспериментальных и теоретических значений Y

X	Экспериментальные значения Y	Теоретические значения Y

0,1	5,998	5,9364
0,2	5,82	5,8788
0,3	5,754	5,8212
0,4	5,828	5,7636

5. Построить на одной координатной плоскости два графика зависимости y от x - экспериментальный и полученный в результате линейного приближения, как показано на рисунке 1. Сделать вывод о качестве линейного приближения.

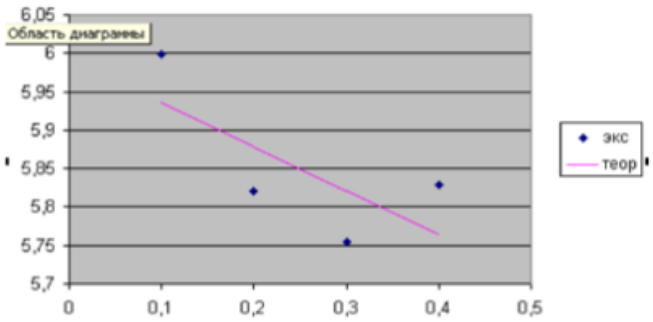


Рисунок 1. Графики зависимости y от x – экспериментальный и полученный в результате линейного приближения

6. Спрогнозируйте значение функции Y еще на два шага вперед от максимального *X*.

Аппроксимация экспериментальных данных в MS Excel

Цель работы. Исследовать принцип проведения аппроксимации экспериментальных данных.

Краткие теоретические сведения.

Задача аппроксимации распадается на две составляющие. Сначала устанавливают вид зависимости $y = f(x)$ и, соответственно, вид эмпирической формулы, то есть решают, является ли она линейной, квадратичной, логарифмической или какой-либо другой. После этого определяются численные значения неизвестных пара метров выбранной эмпирической формулы, для которых приближение к заданной функции оказывается наилучшим. Если нет каких-либо теоретических соображений для подбора вида формулы, обычно выбирают функциональную зависимость из числа наиболее простых, сравнивая их графики с графиком заданной функции.

После выбора вида формулы определяют ее параметры. Для наилучшего выбора параметров задают меру близости аппроксимации экспериментальных данных. Во многих случаях, в особенности если функция f(x) задана графиком или таблицей (на дискретном множестве точек), для оценки степени приближения рассматривают разности $f(xi) - \varphi(xi)$ для точек $x_0, x_1, \ldots, x_n$.

Существуют различные меры близости и, соответственно, способы решения этой задачи. Некоторые из них очень просты, быстро приводят к результату, но результат этот является сильно приближенным. Другие более точные, но и более сложные. Обычно оценку определения параметров при известном виде зависимости осуществляют по методу наименьших квадратов. При этом функция $\varphi(x)$ считается наилучшим приближением к $f(x)$, если для нее сумма квадратов отклонений «теоретических» значений $\varphi(x_i)$, найденных по эмпирической формуле, от соответствующих опытных значений y_i:

$$Z = \sum_{i=0}^{n} [f(x_i) - \varphi(x_i)]^2 \to min \quad (\mathbf{3})$$

Метод наименьших квадратов формулирует аналитические условия достижения суммой квадратов отклонений (3) своего наименьшего значения.

В простейшем случае задача аппроксимации экспериментальных данных выглядит следующим образом.

Пусть есть какие-то данные, полученные практическим путем (в ходе эксперимента или наблюдения), которые можно представить парами чисел (x; y).

Зависимость между ними отражает следующая таблица на рисунке 2.

x	x_1	…	x_n
y	y_1	…	y_n

Рисунок 2. Зависимость между парами чисел

На основе этих данных требуется подобрать функцию $y = \varphi(x)$, которая наилучшим образом сглаживала бы экспериментальную зависимость между переменными и по возможности точно отражала общую тенденцию зависимости между x и y, исключая погрешности измерений и случайные отклонения. Это значит, что отклонения $y_i - y_i(x_i)$ должны быть наименьшими.

Выяснить вид функции можно либо из теоретических соображений, либо

анализируя расположение точек $(x_i; y_i)$ на координатной плоскости.

Например, пусть точки расположены так, как показано на рисунке 3.

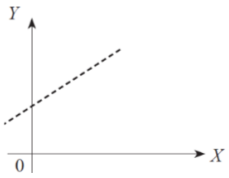


Рисунок 3. Возможный вариант расположения экспериментальных точек

Учитывая то, что практические данные получены с некоторой погрешностью, обусловленной неточностью измерений, необходимостью округления результатов и т. п., естественно предположить, что здесь имеет место линейная зависимость $y=ax+b$.

Чтобы функция приняла конкретный вид, необходимо вычислить a и b.

Построение эмпирической функции сводится к вычислению входящих в нее параметров так, чтобы из всех функций такого вида выбрать ту, которая лучше других описывает зависимость между изучаемыми величинами. То есть сумма квадратов разности между табличными значениями функции в некоторых точках и значениями, вычисленными по полученной формуле, должна быть минимальна.

В MS Excel аппроксимация экспериментальных данных осуществляется путем построения их графика (х – отвлеченные величины) или точечного графика (х – имеет конкретные значения) с последующим подбором подходящей аппроксимирующей функции (линии тренда). Возможны следующие варианты функций:

1 Линейная: $y=ax+b$. Обычно применяется в простейших случаях, когда экспериментальные данные возрастают или убывают с постоянной скоростью.

2 Полиномиальная: $y = a_0 + a_1x + a_2x^2 + ... + a_nx^n$ до шестого порядка .включительно ($n \leq 6$), a_i – константы. Используется для описания экспериментальных данных, попеременно возрастающих и убывающих. Степень полинома определяется количеством экстремумов (максимумов или минимумов).

Полином второй степени может описать только один максимум или минимум, полином третьей степени может иметь один или два экстремума, четвертой степени – не более трех экстремумов и т. д.

3 Логарифмическая: $y = a\ lnx + b$, где a и b – константы, lnx – функция натурального логарифма. Функция применяется для описания экспериментальных данных, которые вначале быстро растут или убывают, а затем постепенно стабилизируются.

4 Степенная: $y = bx^a$, где a и b – константы. Аппроксимация степенной функцией используется для экспериментальных данных с постоянно увеличивающейся (или убывающей) скоростью роста. Данные не должны иметь нулевых или отрицательных значений.

5 Экспоненциальная: $y = b \cdot e^{ax}$, где a и b – константы, e – основание натурального логарифма. Применяется для описания экспериментальных данных, которые быстро растут или убывают, а затем постепенно стабилизируются. Часто ее использование вытекает из теоретических соображений.

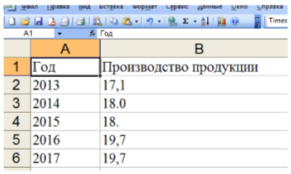
Степень близости аппроксимации экспериментальных данных выбранной функцией оценивается коэффициентом детерминации (R_2). Таким образом, если есть несколько подходящих вариантов типов аппроксимирующих функций, можно выбрать функцию с большим коэффициентом детерминации (стремящимся к 1).

Для осуществления аппроксимации на диаграмме экспериментальных данных в случае использования пакета Microsoft Excel необходимо щелчком правой кнопки мыши вызвать контекстное меню и выбрать пункт Добавить линию тренда.

В появившемся диалоговом окне Линия тренда на вкладке Тип выбирается вид аппроксимирующей функции, а на вкладке Параметры задаются дополнительные параметры, влияющие на отображение аппроксимирующей кривой [1].

Задание на выполнение лабораторной работы.

Исследовать характер изменения с течением времени уровня производства некоторой продукции и подобрать аппроксимирующую функцию, располагая следующими данными, показанными на рисунке 4.



	A	B
1	Год	Производство продукции
2	2013	17,1
3	2014	18.0
4	2015	18.
5	2016	19,7
6	2017	19,7

Рисунок 4. Данные производства некоторой продукции

Порядок выполнения работы:

1 Для построения диаграммы прежде всего необходимо ввести данные в рабочую таблицу.

2 Далее по введенным в рабочую таблицу данным необходимо построить диаграмму, как показано на рисунке 5. Поскольку здесь необходимо показать динамику изменений производства продукции, не привязываясь к конкретному году, а от отвлеченных переменных, выберем диаграмму График.

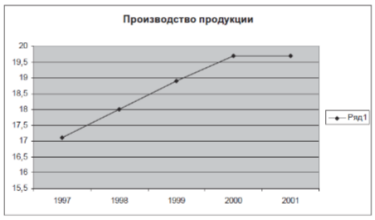


Рисунок 5. График экспериментальных данных

3 Осуществим аппроксимацию полученной кривой полиномиальной функцией второго порядка, поскольку кривая довольно гладкая и не сильно отличается от прямой линии. Для этого указатель мыши устанавливаем на одну из точек графика и щелкаем правой кнопкой. В появившемся контекстном меню выбираем пункт. Добавить линию тренда. Появляется диалоговое окно Линия тренда, как показано на рисунке 6.

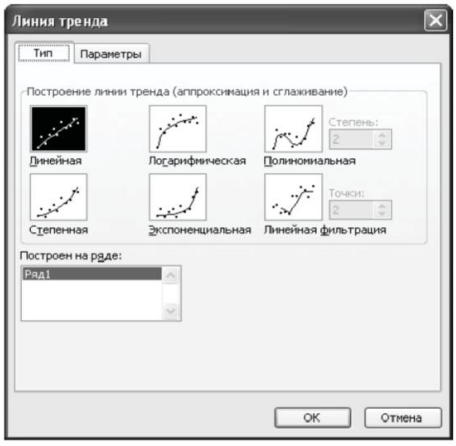


Рисунок 6. Вкладка Тип диалогового окна **Линия тренда**

В этом окне на вкладке Тип выбираем тип линии тренда – Полиномиальная – и устанавливаем степень – 2. Затем открываем вкладку Параметры, показанный на рисунке 7 и устанавливаем флажки в поля показывать уравнение на диаграмме и поместить на диаграмму величину достоверности аппроксимации (R_2).

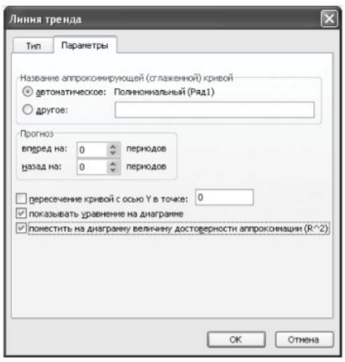


Рисунок 7. Вкладка Параметры диалогового окна **Линия тренда**

После чего нужно щелкнуть по кнопке ОК. В результате получим на диаграмме аппроксимирующую кривую, показанную на рисунке 8.

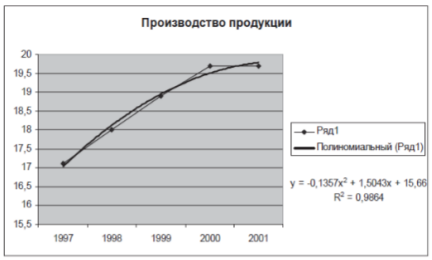


Рисунок 8. Экспериментальные данные, аппроксимированные полиномиальной кривой

Как видно из рисунка 8, уравнение наилучшей полиномиальной аппроксимирующей функции для некоторых отвлеченных значений x (1, 2, 3, ...) выглядит как $y = -0,14x^2 +1,5x + 15,66$.

При этом точность аппроксимации достаточно высока – $R_2 = 0,986$.