ЛАБОРАТОРНАЯ РАБОТА 2.
ЛИНЕЙНАЯ РЕГРЕССИЯ. КРИВОЛИНЕЙНАЯ РЕГРЕССИЯ

Задание. Конденсатор заряжен до напряжения U_0, отвечающего моменту начала отсчета времени, после чего он разряжается через некоторое сопротивление. Напряжение измеряется с округлением до 5 В. Исследовать зависимость напряжения U от времени t. Основные результаты и выводы по пунктам оформить письменно.

t	0	1	2	3	4	5	6	7	8	9	10
U	95	75	55	40	30	20	15	10	10	5	5

1. Построить корреляционное поле.
2. Вычислить выборочный коэффициент корреляции, проверить его значимость на уровне значимости $\alpha = 0{,}05$.
3. По характеру расположения точек на корреляционном поле и на основании проверки значимости коэффициента корреляции сделать вывод о соответствии или несоответствии линейной модели экспериментальным данным.
4. Составить систему нормальных уравнений для определения по методу наименьших квадратов коэффициентов линейного уравнения регрессии, найти выборочное уравнение линейной регрессии, построить прямую на корреляционном поле.
5. Подтвердить либо опровергнуть вывод пункта 3.
6. С помощью Мастера диаграмм в Excel получить (если это возможно) уравнения следующих зависимостей:

   а) $y = b_0 + b_1 x$; б) $y = b_0 + b_1 x + b_2 x^2$; в) $y = b_0 + \dfrac{b_1}{x}$;

   г) $y = b_0 + b_1 \ln x$; д) $y = a\,\mathrm{e}^{bx}$; е) $y = ax^b$; ж) $y = \dfrac{1}{b_0 + b_1 x}$.

   Указание 1. Если все значения переменной y отрицательны, для получения зависимостей д) и е) следует сделать замену $Y = |y|$.

   Указание 2. Для получения гиперболических зависимостей в) и ж) нужно построить линейные зависимости на новых диаграммах, сделав соответствующие замены переменных.
7. Сравнить уравнение а) с полученным в пункте 4.
8. На основании значений коэффициента детерминации R^2 сделать вывод о наилучшей модели из допустимых.
9. *В случае б): составить систему нормальных уравнений для определения по методу наименьших квадратов коэффициентов квадратичного уравнения регрессии; найти выборочное квадратичное уравнение регрессии.

   В случаях в)-ж): указать замену переменных, позволяющую свести выбранную зависимость к линейной; построить корреляционное поле в новых переменных; составить систему нормальных уравнений для определения по методу наименьших квадратов коэффициентов линейного уравнения регрессии

в новых переменных; найти выборочное уравнение линейной регрессии, построить прямую на корреляционном поле; сделав обратную замену, получить уравнение регрессии в натуральных переменных.

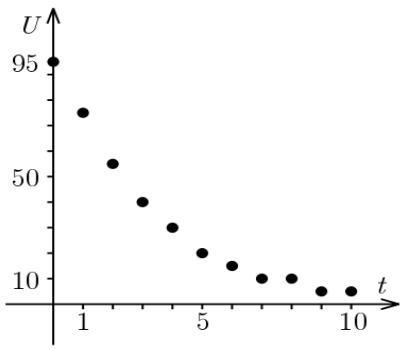
Контрольные вопросы

1. Виды зависимостей между двумя СВ.
2. В чем различие между статистической и функциональной зависимостями двух СВ?
3. Что такое регрессионная зависимость между двумя СВ?
4. Основные задачи корреляционного анализа.
5. Основные задачи регрессионного анализа.
6. На основании чего осуществляется выбор вида функции регрессии?
7. Что называется корреляционным полем?
8. Почему наиболее часто используется модель линейной регрессии?
9. Какой статистический показатель используется в качестве количественной меры линейной связи между двумя наблюдаемыми величинами?
10. Свойства выборочного коэффициента корреляции.
11. Какие значения может принимать выборочный коэффициент корреляции?
12. Какие значения принимает выборочный коэффициент корреляции, если наблюдаемые величины независимы?
13. Какие значения принимает выборочный коэффициент корреляции, если наблюдаемые величины связаны линейной зависимостью?
14. Что показывает знак выборочного коэффициента корреляции?
15. Для чего проводится проверка значимости коэффициента корреляции?
16. Как проводится проверка значимости коэффициента корреляции в случае, если наблюдаемые величины имеют совместное нормальное распределение?
17. В чем суть метода наименьших квадратов?
18. Система нормальных уравнений метода наименьших квадратов.
19. Как связан коэффициент детерминации с коэффициентом корреляции в случае линейной регрессионной модели?
20. С помощью какой замены переменных можно свести к линейной следующие зависимости: а) $y = b_0 + \frac{b_1}{x}$; б) $y = b_0 + b_1 \ln x$; в) $y = a\,\mathrm{e}^{bx}$; г) $y = ax^b$; д) $y = \frac{1}{b_0 + b_1 x}$?

Пример и методические указания по выполнению лабораторной работы

1. По результатам $n = 11$ измерений исследуем зависимость напряжения U от времени t.

Построим корреляционное поле. По виду корреляционного поля можно предположить, что *выборочный коэффициент корреляции отрицателен и значимо отличается от 0.* (Почему?)



2. Обозначим через x независимую переменную t (время), через y – зависимую переменную U (напряжение).

Расчетная таблица

	x_i	y_i	x_iy_i	x_i^2	y_i^2
1	0	95	0	0	9025
2	1	75	75	1	5625
3	2	55	110	4	3025
4	3	40	120	9	1600
5	4	30	120	16	900
6	5	20	100	25	400
7	6	15	90	36	225
8	7	10	70	49	100
9	8	10	80	64	100
10	9	5	45	81	25
11	10	5	50	100	25
$\sum$	**55**	**360**	**860**	**385**	**21050**

Выборочный коэффициент корреляции вычислим по формуле

$$r_{x;\,y} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\sqrt{D_{\text{в}}(x)D_{\text{в}}(y)}},$$

где

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{55}{11} = 5; \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{360}{11} \approx 32,7;$$

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \frac{860}{11} \approx 78,2;$$

$$D_{\text{в}}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2 = \frac{385}{11} - 5^2 = 10;$$

$$D_{\text{в}}(y) = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - (\bar{y})^2 = \frac{21050}{11} - 32,7^2 \approx 844,35.$$

Тогда

$$r_{x;\,y} = \frac{78,2 - 5\cdot 32,7}{\sqrt{10\cdot 844,35}} \approx -0,928.$$

Что можно сказать о зависимости между величинами, если выборочный коэффициент корреляции отрицательный?

Проверка значимости коэффициента корреляции. Вычислим расчетное значение критерия Стьюдента:

$$t_{\text{расч}} = \left|r_{x;\,y}\right|\sqrt{\frac{n-2}{1-r_{x;\,y}^2}} = 0,928\sqrt{\frac{11-2}{1-0,928^2}} \approx 7,47 > t_{\text{табл}} = t_{\alpha;\,n-2},$$

и найдем по таблице квантилей распределения Стьюдента

$$t_{\text{табл}} = t_{0,05;\,9} \approx \frac{2,23+2,31}{2} = 2,27.$$

Поскольку $t_{\text{расч}} = 7,47 > t_{\text{табл}} = 2,27,$ то ***при уровне значимости α = 0,05 коэффициент корреляции считаем значимо отличающимся от нуля***, а следовательно, связь между величинами x, y признается статистически значимой, т.е. *результаты исследований не случайны и могут быть признаны достоверными*.

3. Поскольку коэффициент корреляции признается значимо отличающимся от нуля, можно принять предположение о линейной регрессионной зависимости между наблюдаемыми величинами. Однако расположение точек на корреляционном поле свидетельствует о другой, криволинейной зависимости.

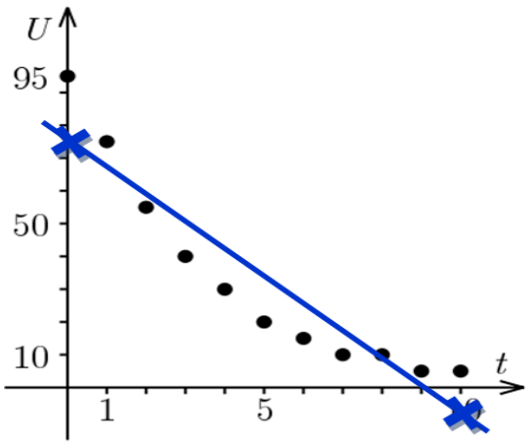
4. Система нормальных уравнений для определения МНК-коэффициентов b_0 и b_1 линейного эмпирического уравнения регрессии $\hat{y} = b_0 + b_1 x.$

$$\begin{cases} nb_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \\ b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \end{cases} \qquad \begin{cases} 11b_0 + 55b_1 = 360, \\ 55b_0 + 385b_1 = 860. \end{cases}$$

Какими методами можно решить систему линейных алгебраических уравнений?

Итак, эмпирическое линейное уравнение регрессии имеет вид $\hat{y} = 75,45 - 8,55x.$

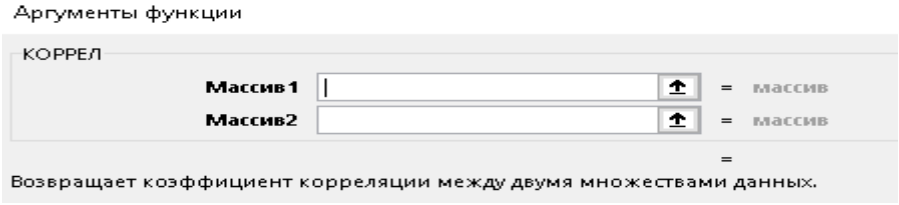
Прямая на корреляционном поле:
если $x = 0$, то $\hat{y} = 75,45$;
если $x = 10$, то $\hat{y} = -10,05.$

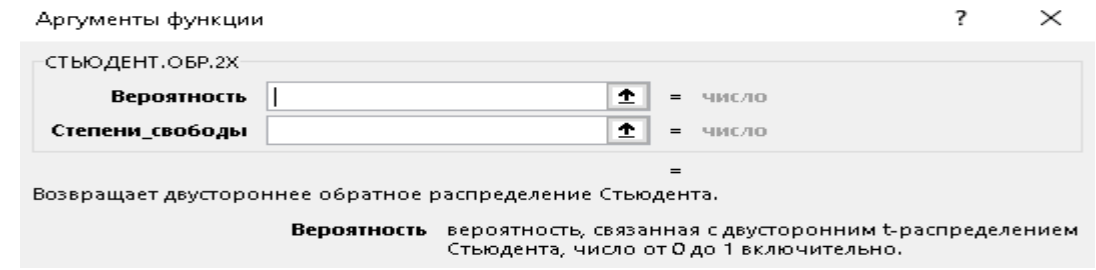


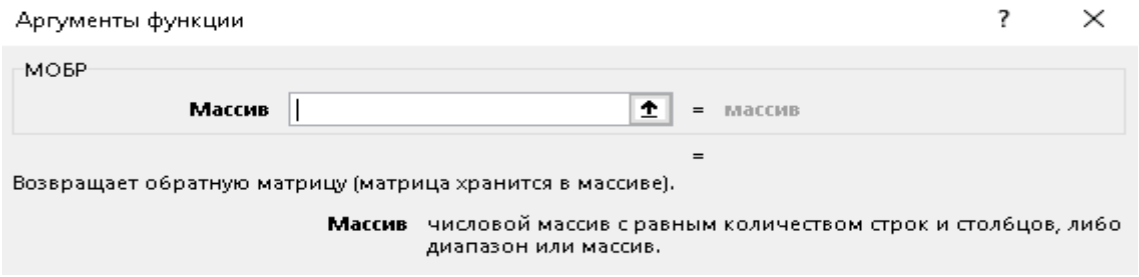
Согласно МНК, **построенная прямая приближает экспериментальные данные наилучшим образом. Что это означает?**

5. Подтверждаем вывод пункта 3 о том, что полученная прямая удовлетворительно приближает экспериментальные данные, однако расположение экспериментальных точек свидетельствует о наличии другой, криволинейной зависимости между наблюдаемыми величинами.

При выполнении пунктов 1-5 в Excel можно использовать встроенные функции



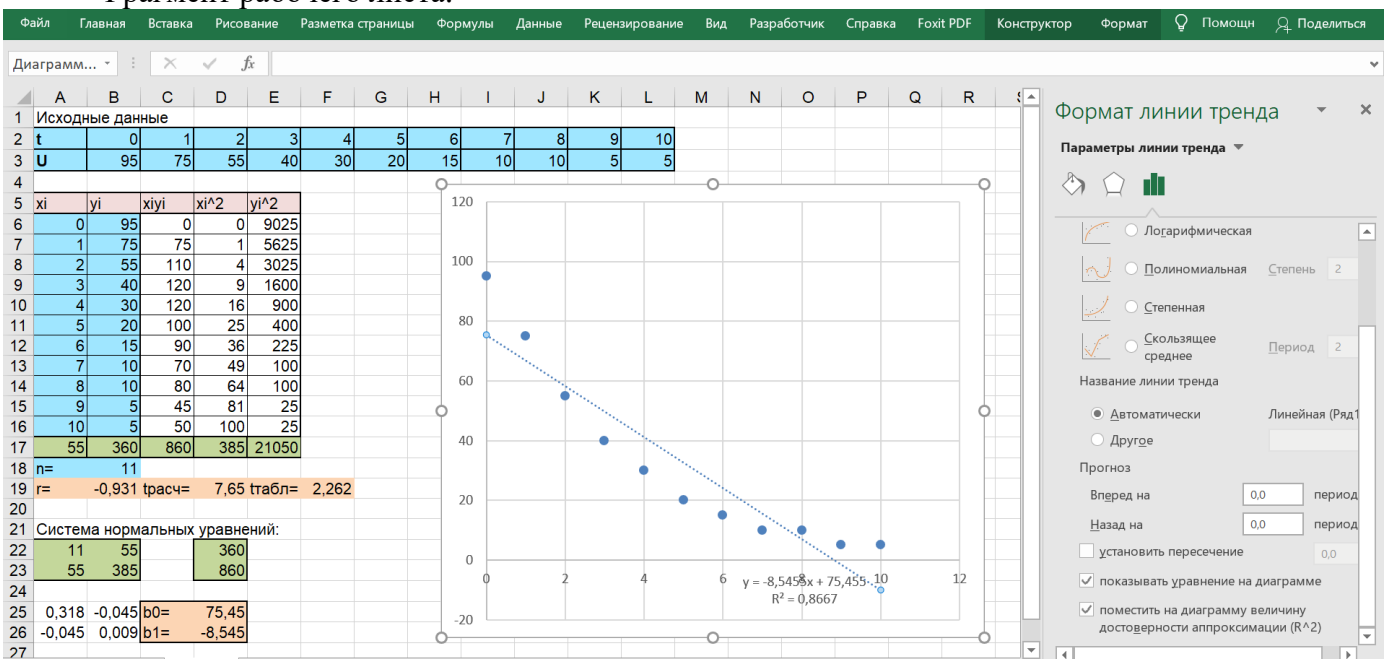




Выделяем массив (в который будет записана обратная матрица), вызываем функцию =МОБР и нажимаем сочетание клавиш ***Ctrl+Shift+Enter*** *(три клавиши вместе!).*

Для построения корреляционного поля выделите массив данных и выберите *Вставка → Диаграммы → Точечная.*

Фрагмент рабочего листа.

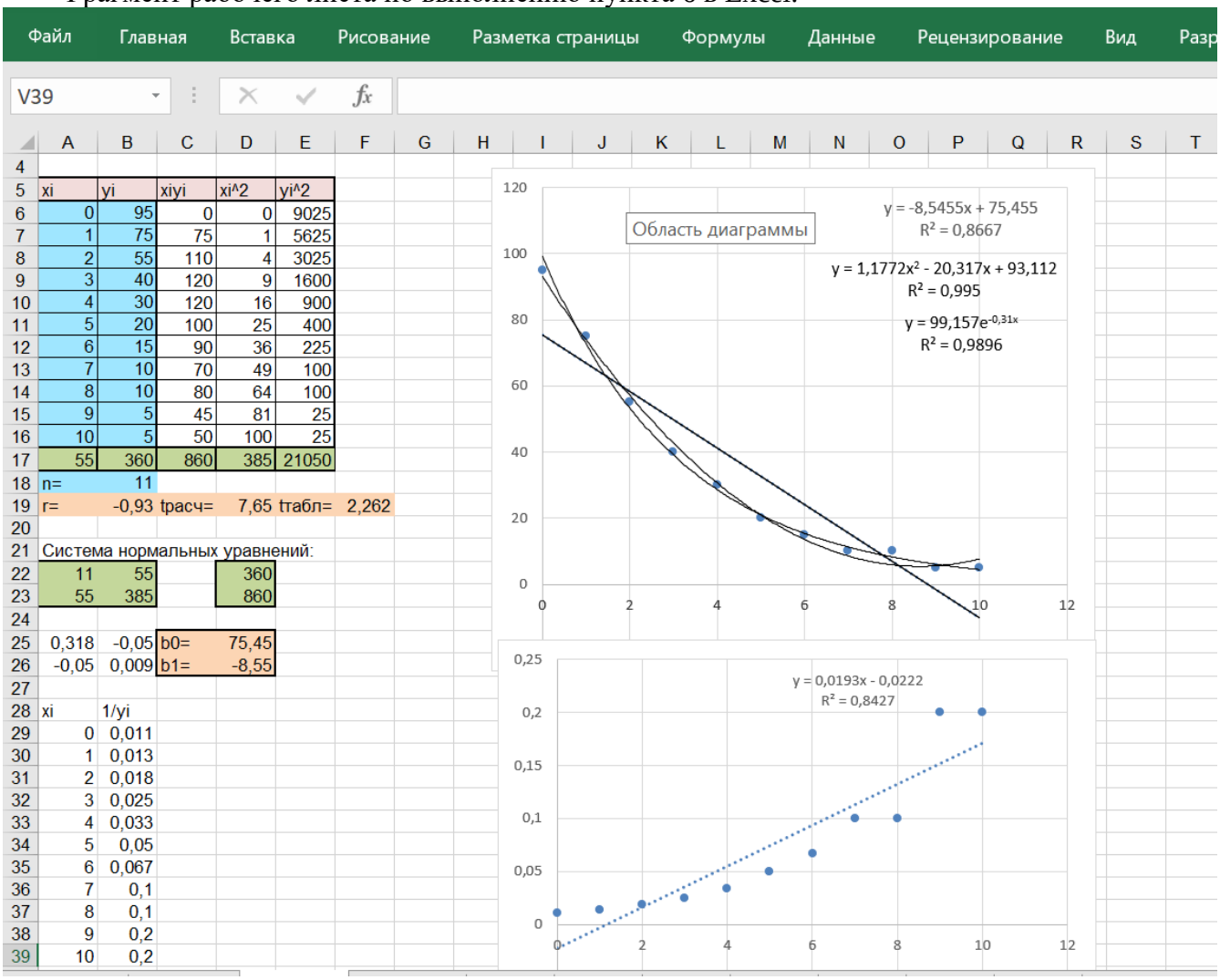


6. С помощью Excel подберем наилучшую аппроксимирующую функцию для исходных данных. **Заполните таблицу в письменном отчете в соответствии с решением Вашего варианта**

Вид зависимости	Уравнение зависимости	Коэффициент детерминации R^2	Примечание
а) Линейная	$y = -8,5455x + 75,455$	0,8667	
б) Квадратичная	$y = 1,1772x^2 - 20,317x + 93,112$	0,995	
в) Гиперболическая	–	–	есть значение $x = 0$
г) Логарифмическая	–	–	есть значение $x = 0$
д) Экспоненциальная	$y = 99,157\,e^{-0,31x}$	0,9896	
е) Степенная	–	–	есть значение $x = 0$

ж) Гиперболическая	$y = \dfrac{1}{0{,}0193x - 0{,}0222}$	0,8427	

Фрагмент рабочего листа по выполнению пункта 6 в Excel.



	A	B	C	D	E
5	xi	yi	xiyi	xi^2	yi^2
6	0	95	0	0	9025
7	1	75	75	1	5625
8	2	55	110	4	3025
9	3	40	120	9	1600
10	4	30	120	16	900
11	5	20	100	25	400
12	6	15	90	36	225
13	7	10	70	49	100
14	8	10	80	64	100
15	9	5	45	81	25
16	10	5	50	100	25
17	55	360	860	385	21050
18	n=	11			
19	r=	-0,93	tрасч=	7,65	tтабл= 2,262
21	Система нормальных уравнений:				
22	11	55		360	
23	55	385		860	
25	0,318	-0,05	b0=	75,45	
26	-0,05	0,009	b1=	-8,55	
28	xi	1/yi			
29	0	0,011			
30	1	0,013			
31	2	0,018			
32	3	0,025			
33	4	0,033			
34	5	0,05			
35	6	0,067			
36	7	0,1			
37	8	0,1			
38	9	0,2			
39	10	0,2			

Чтобы получить на диаграмме уравнение регрессии, щелкните правой кнопкой мыши по одной из точек на диаграмме и выберите Добавить линию тренда..., укажите тип линии тренда и поставьте две галочки:

- ✓ *показывать уравнение на диаграмме*
- ✓ *поместить на диаграмму величину достоверности аппроксимации (R^2)*

7. Полученное в расчетах пункта 4 уравнение регрессии $\hat{y} = 75{,}45 - 8{,}55x$ совпадает с уравнением линейной линии тренда $y = -8{,}5455x + 75{,}455$; квадрат коэффициента корреляции $r^2_{x;y} = (-0{,}928)^2 \approx 0{,}861$ приблизительно равен коэффициенту детерминации R^2 = 0,8667 (различие объясняется округлениями при вычислениях в пункте 4).

8. Выберем из полученных уравнений наилучшую аппроксимирующую функцию, учитывая значения коэффициента детерминации R^2 и сложность модели.

Наибольший коэффициент детерминации R^2 имеет квадратичная зависимость, однако это значение R^2 = 0,995 незначительно превышает значение $R^2 = 0{,}9896$ для экспоненциальной модели, которая проще в том смысле, что содержит меньше параметров (коэффициентов).

Вид корреляционного поля (точки группируются вдоль убывающей кривой, вторая ветвь параболы не прослеживается) и физическая сущность данных (напряжение с течением времени должно уменьшаться и стремиться к нулю) свидетельствуют в пользу экспоненциальной модели.

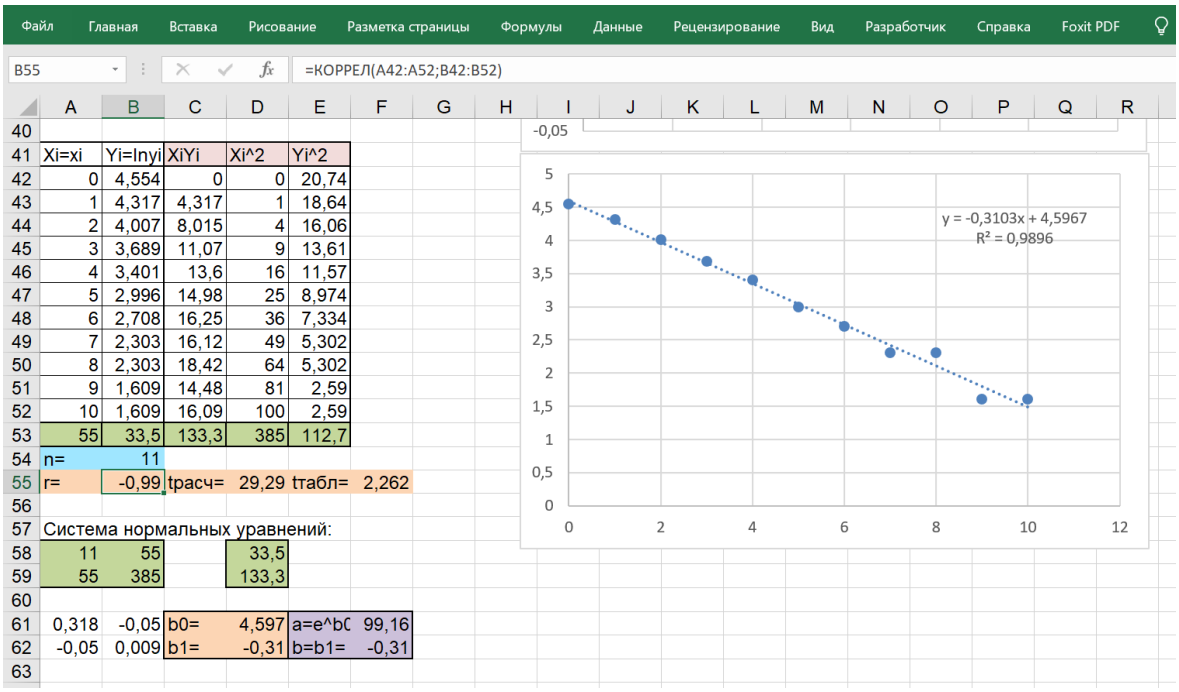
Таким образом, наилучшей аппроксимирующей функцией признаем экспоненциальную функцию $y = 99{,}157\,\mathrm{e}^{-0{,}31x}$ с $R^2 = 0{,}9896$.

9. * Параметры экспоненциальной зависимости $y = a\,\mathrm{e}^{bx}$ могут быть получены с помощью МНК, поскольку эта зависимость может быть сведена к линейной с помощью логарифмирования:

$$\ln y = \ln a + \ln \mathrm{e}^{bx} \quad \Rightarrow \quad \ln y = \ln a + bx.$$

Если ввести новые переменные $Y = \ln y,\ X = x,$ исходная зависимость сведется к линейной $Y = b_0 + b_1 X,$ коэффициенты которой могут быть найдены по МНК. Тогда коэффициенты искомой зависимости определятся из соотношений $a = \mathrm{e}^{b_0},\ b = b_1.$

Для проверки того, удачно ли выбран вид зависимости, построим новое корреляционное поле на плоскости *OXY* (см. фрагмент рабочего листа Excel).



	A	B	C	D	E
41	Xi=xi	Yi=lnyi	XiYi	Xi^2	Yi^2
42	0	4,554	0	0	20,74
43	1	4,317	4,317	1	18,64
44	2	4,007	8,015	4	16,06
45	3	3,689	11,07	9	13,61
46	4	3,401	13,6	16	11,57
47	5	2,996	14,98	25	8,974
48	6	2,708	16,25	36	7,334
49	7	2,303	16,12	49	5,302
50	8	2,303	18,42	64	5,302
51	9	1,609	14,48	81	2,59
52	10	1,609	16,09	100	2,59
53	55	33,5	133,3	385	112,7
54	n=	11			
55	r=	-0,99	tрасч=	29,29	tтабл= 2,262
56					
57	Система нормальных уравнений:				
58	11	55		33,5	
59	55	385		133,3	
60					
61	0,318	-0,05	b0=	4,597	a=e^b0 99,16
62	-0,05	0,009	b1=	-0,31	b=b1= -0,31
63					

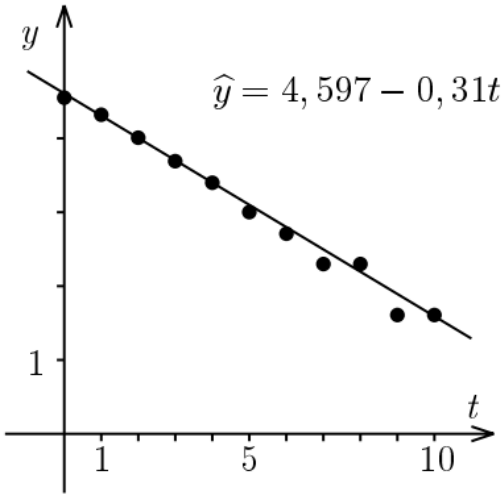
На диаграмме точки $(X_i; Y_i)$ располагаются вдоль прямой, коэффициент корреляции $r_{X;Y} = -0{,}99,$ а значит, вид зависимости *y* от *x* подобран правильно.

Коэффициенты линейного уравнения регрессии $Y = b_0 + b_1 X$ в новых переменных найдем из системы нормальных уравнений МНК:

$$\begin{cases} n b_0 + b_1 \sum\limits_{i=1}^{n} X_i = \sum\limits_{i=1}^{n} Y_i, \\ b_0 \sum\limits_{i=1}^{n} X_i + b_1 \sum\limits_{i=1}^{n} X_i^2 = \sum\limits_{i=1}^{n} X_i Y_i; \end{cases} \qquad \begin{cases} 11 b_0 + 55 b_1 = 33{,}5, \\ 55 b_0 + 385 b_1 = 133{,}3. \end{cases}$$

Решая систему матричным методом, получим:

$$b_0 = 4{,}597,\ b_1 = -0{,}31 \quad \Rightarrow \quad Y = 4{,}597 - 0{,}31X.$$



Следовательно,

$$a = \mathrm{e}^{b_0} = \mathrm{e}^{4{,}597} = 99{,}16,\ b = b_1 = -0{,}31 \quad \Rightarrow \quad y = 99{,}16\,\mathrm{e}^{-0{,}31x},$$

что совпадает с уравнением экспоненциальной линии тренда, полученным в пункте 6.

Замечание. В таблице указаны преобразования, с помощью которых можно «выровнять» некоторые зависимости, наиболее часто встречающиеся на практике.

Вид зависимости	Уравнение зависимости	Замена переменных, сводящая зависимость к линейной $Y = b_0 + b_1 X$	Выражение параметров зависимости через коэффициенты b_0, b_1
Гиперболическая	$y = a + \frac{b}{x}$	$Y = y,\ X = \frac{1}{x}$	$a = b_0,\ b = b_1$
Логарифмическая	$y = a + b\ln x$	$Y = y,\ X = \ln x$	$a = b_0,\ b = b_1$
Экспоненциальная	$y = a\,\mathrm{e}^{bx}$	$Y = \ln y,\ X = x$	$a = \mathrm{e}^{b_0},\ b = b_1$
Степенная	$y = ax^b$	$Y = \ln y,\ X = \ln x$	$a = \mathrm{e}^{b_0},\ b = b_1$
Гиперболическая	$y = \frac{1}{a + bx}$	$Y = \frac{1}{y},\ X = x$	$a = b_0,\ b = b_1$