Лабораторная работа 2.
Линейная регрессия. Криволинейная регрессия

Теоретические сведения

Пусть изучается связь между двумя величинами на основании экспериментальных данных – по выборке объема n пар значений: $(x_1; y_1)$, $(x_2; y_2)$, ..., $(x_n; y_n)$.

Две случайные величины (СВ) могут быть: 1) независимыми; 2) связаны функциональной зависимостью (каждому значению одной из них соответствует строго определенное значение другой); 3) связаны статистической зависимостью (каждому значению одной СВ соответствует множество возможных значений другой и изменение значения одной величины влечет изменение *распределения* другой).

При изучении статистической зависимости обычно ограничиваются исследованием усредненной зависимости: как в среднем будет изменяться значение одной величины при изменении другой. Такая зависимость называется ***регрессионной***.

Основным методом исследования статистических зависимостей является ***корреляционно-регрессионный*** анализ.

Основными задачами корреляционного анализа являются выявление связи между наблюдаемыми СВ и оценка тесноты этой связи.

Основными задачами регрессионного анализа являются установление *формы зависимости* между наблюдаемыми величинами и определение по экспериментальным данным уравнения зависимости, которое называют ***выборочным (эмпирическим) уравнением регрессии***, а также прогнозирование с помощью уравнения регрессии среднего значения зависимой переменной при заданном значении независимой переменной.

Вид эмпирической функции регрессии определяют исходя из: 1) соображений о физической сущности исследуемой зависимости; 2) опыта предыдущих исследований; 3) характера расположения точек на ***корреляционном поле***, которое получается, если отметить на плоскости все точки с координатами (x_i, y_i), соответствующие наблюдениям.

Наибольший интерес представляет линейное эмпирическое уравнение регрессии $\hat{y} = b_0 + b_1 x$, так как: 1) это наиболее простой случай для расчетов и анализа; 2) при нормальном распределении функция регрессии является линейной.

Количественной мерой *линейной связи* между двумя наблюдаемыми величинами служит ***выборочный коэффициент корреляции***.

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

где $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$ – выборочные средние величин x, y и произведения xy соответственно; $D_{\text{в}}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2$, $D_{\text{в}}(y) = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - (\bar{y})^2$ – выборочные дисперсии величин x и y.

Свойства выборочного коэффициента корреляции.

1. $-1 \le r_{x;\,y} \le 1$.

2. Если наблюдаемые величины x и y независимы, то $r_{x;\,y} \approx 0$. Однако обратное неверно: значение $r_{x;\,y} \approx 0$ не гарантирует, что наблюдаемые величины x и y независимы.

3. Если $\left|r_{x;\,y}\right| = 1$ (или близок к 1), то наблюдаемые величины x и y связаны линейной зависимостью, т. е. $y = b_0 + b_1 x$.

4. Если $r_{x;\,y} > 0,$ то с ростом значений одной величины значения другой также в основном возрастают; если $r_{x;\,y} < 0,$ то с ростом значений одной величины значения другой, наоборот, убывают.

Проверка значимости коэффициента корреляции – это проверка гипотезы о том, что коэффициент корреляции значимо отличается от нуля. Так как выборка произведена случайно, нельзя утверждать, что если выборочный коэффициент корреляции $r_{x;\,y} \neq 0,$ то и коэффициент корреляции генеральной совокупности $r_{\xi;\,\eta} \neq 0$. Возможно, отличие $r_{x;\,y}$ от 0 вызвано только случайными искажениями наблюдаемых значений.

Если выборка из нормального распределения, то проверка производится по *критерию Стьюдента*: если

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

где $t_{\alpha;\,n-2}$ – квантиль уровня α распределения Стьюдента с числом степеней свободы $k = n - 2$ (определяется по таблице), то при заданном уровне значимости α (допускается, что вывод может быть ошибочным с небольшой вероятностью α) коэффициент корреляции считается значимо отличающимся от нуля, а следовательно, связь между величинами x, y признается статистически значимой.

Если коэффициент корреляции на основании проверки признается значимо отличающимся от нуля, считают допустимым принять предположение о линейной регрессионной зависимости между наблюдаемыми величинами.

Подчеркнем, что *коэффициент корреляции является мерой именно линейной зависимости*. В случае нелинейной зависимости связь между величиной коэффициента корреляции и близостью точек корреляционного поля к некоторой линии не прослеживается. Поэтому в практических задачах при выборе вида эмпирической функции регрессии обязательно учитывают характер расположения точек на корреляционном поле.

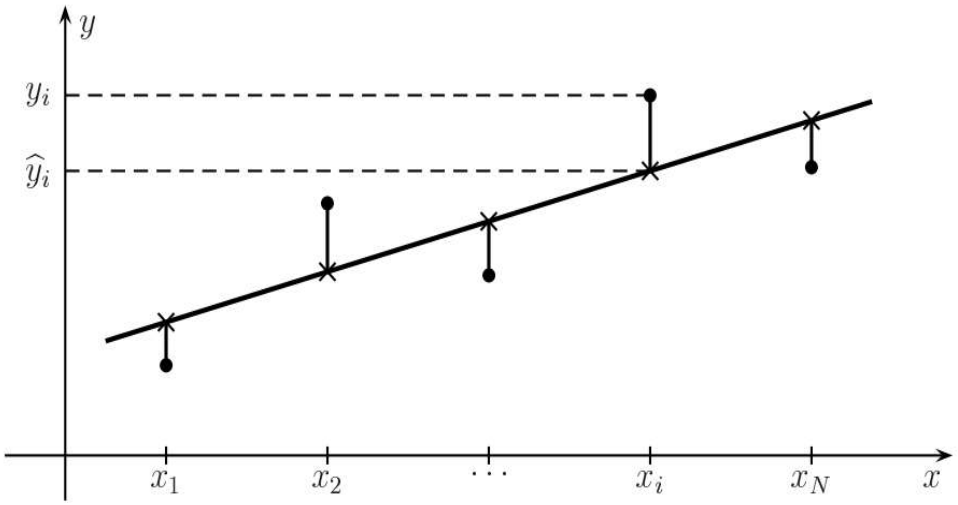
Определение коэффициентов эмпирического линейного уравнения регрессии методом наименьших квадратов. Пусть имеется выборка объема n наблюдений над двумя величинами x и y: $(x_1; y_1)$, $(x_2; y_2)$, ..., $(x_n; y_n)$, и принята гипотеза о линейной зависимости между y и x. Для определения коэффициентов линейного эмпирического уравнения регрессии

$$\hat{y} = b_0 + b_1 x$$

используется ***метод наименьших квадратов*** (***МНК***). Суть этого метода в том, что коэффициенты b_0 и b_1 выбирают так, чтобы сумма квадратов отклонений

наблюдаемых значений y_i от предсказываемых по уравнению $\hat{y}_i = b_0 + b_1 x_i$ была минимальной (см. рис.). Таким образом, минимизируется функция

$$Q(b_0; b_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \to \min_{b_0, b_1}.$$



Согласно МНК, значения параметров b_0 и b_1 находят из системы, которая называется ***системой нормальных уравнений*** метода наименьших квадратов:

$$\begin{cases} nb_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \\ b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \end{cases}$$

Метод наименьших квадратов широко применяется при статистической обработке результатов измерений.

Зависимость между двумя наблюдаемыми величинами далеко не всегда можно выразить линейной функцией. Иногда видно, что точки корреляционного поля образуют некоторую кривую. При выборе вида эмпирической функции регрессии необходимо учитывать теоретические сведения и опыт предыдущих аналогичных исследований.

Как правило, до начала исследования должен быть определен вид эмпирической функции регрессии с точностью до нескольких параметров, значения которых оцениваются по результатам эксперимента. В том случае, если функция регрессии линейна по параметрам или может быть сведена к таковой с помощью замены переменных, для определения оценок параметров используют МНК.

Например, коэффициенты квадратичного уравнения регрессии $\hat{y} = b_0 + b_1 x + b_2 x^2$ находят из следующей системы нормальных уравнений:

$$\begin{cases} nb_0 + b_1\sum\limits_{i=1}^{n} x_i + b_2\sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} y_i, \\ b_0\sum\limits_{i=1}^{n} x_i + b_1\sum\limits_{i=1}^{n} x_i^2 + b_2\sum\limits_{i=1}^{n} x_i^3 = \sum\limits_{i=1}^{n} x_i y_i, \\ b_0\sum\limits_{i=1}^{n} x_i^2 + b_1\sum\limits_{i=1}^{n} x_i^3 + b_2\sum\limits_{i=1}^{n} x_i^4 = \sum\limits_{i=1}^{n} x_i^2 y_i. \end{cases}$$

Степенная зависимость вида $y = ax^b$ может быть сведена к линейной с помощью логарифмирования:

$$\ln y = \ln a + \ln x^b \quad \Rightarrow \quad \ln y = \ln a + b\ln x.$$

Если ввести новые переменные $Y = \ln y, X = \ln x,$ 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 $a = \mathrm{e}^{b_0}, b = b_1.$ В таблице приведены некоторые виды зависимостей, которые сводятся к линейной после замены переменных.

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	$y = \frac{1}{a + bx}$	$Y = \frac{1}{y}, X = x$	$a = b_0, b = b_1$

Для проверки того, удачно ли выбран вид зависимости, следует построить новое корреляционное поле на плоскости *OXY*. Если вид зависимости y от x подобран правильно, то точки $(X_i; Y_i)$ будут располагаться вдоль прямой.

Для выбора наилучшей аппроксимирующей функции из нескольких (в случае, когда нет теоретического обоснования для выбора определенного вида зависимости) используют коэффициент детерминации R^2, который принимает значения от 0 до 1. Чем ближе значение коэффициента к 1, тем сильнее зависимость. В случае линейной зависимости R^2 равен квадрату выборочного коэффициента корреляции.

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 мерой 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 в Excel

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

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

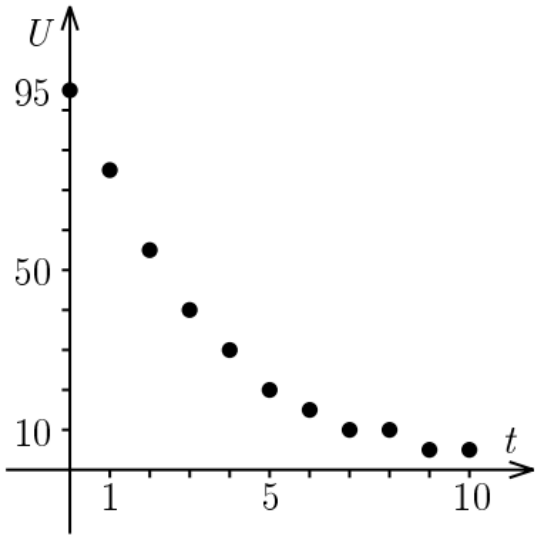
Указание 2. Для получения гиперболических зависимостей в) и ж) нужно построить линейные зависимости на новых диаграммах, сделав соответствующие замены переменных.

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.

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. Требуется определить зависимость напряжения U от времени t.

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. Требуется исследовать зависимость напряжения U от времени t по результатам $n = 11$ измерений. Построим корреляционное поле.



По виду корреляционного поля можно предположить, что выборочный коэффициент корреляции отрицателен и значимо отличается от 0.

2. Для удобства вычислений составим таблицу. Обозначим через x независимую переменную t (время), через y – зависимую переменную U (напряжение). Запишем

исходные данные в столбцы x_i, y_i, добавим столбцы x_iy_i, x_i^2, y_i^2, рассчитаем соответствующие значения и вычислим сумму чисел в каждом столбце.

	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

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

Для проверки 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 $\alpha = 0,05$ коэффициент корреляции считаем значимо отличающимся от нуля, а следовательно, связь между величинами x, y признается статистически значимой.

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. Определим с помощью МНК коэффициенты b_0 и b_1 линейного эмпирического уравнения регрессии $\hat{y} = b_0 + b_1 x.$ Для этого составим систему нормальных уравнений:

$$\begin{cases} nb_0 + b_1 \sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} y_i, \\ b_0 \sum\limits_{i=1}^{n} x_i + b_1 \sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} x_i y_i. \end{cases}$$

Подставляя рассчитанные значения сумм, получим:

$$\begin{cases} 11b_0 + 55b_1 = 360, \\ 55b_0 + 385b_1 = 860. \end{cases}$$

Решим систему по формулам Крамера:

$$b_0 = \frac{\begin{vmatrix} 360 & 55 \\ 860 & 385 \end{vmatrix}}{\begin{vmatrix} 11 & 55 \\ 55 & 385 \end{vmatrix}} = \frac{360 \cdot 385 - 860 \cdot 55}{11 \cdot 385 - 55 \cdot 55} = \frac{91300}{1210} \approx 75,45;$$

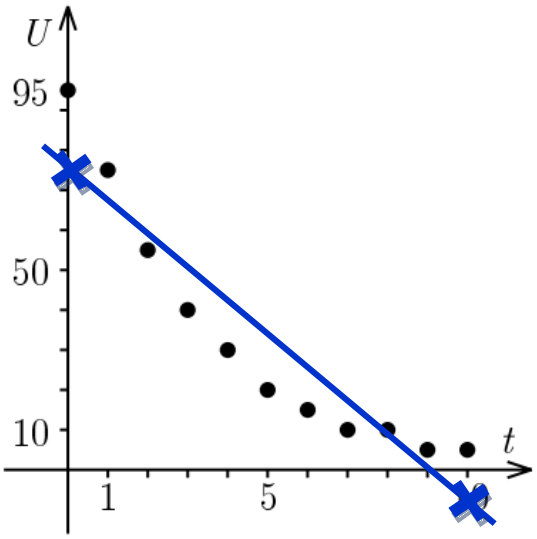
$$b_1 = \frac{\begin{vmatrix} 11 & 360 \\ 55 & 860 \end{vmatrix}}{\begin{vmatrix} 11 & 55 \\ 55 & 385 \end{vmatrix}} = \frac{11 \cdot 860 - 55 \cdot 360}{11 \cdot 385 - 55 \cdot 55} = -\frac{10340}{1210} \approx -8,55.$$

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

Построим прямую на корреляционном поле:

если $x = 0$, то $\hat{y} = 75,45;$

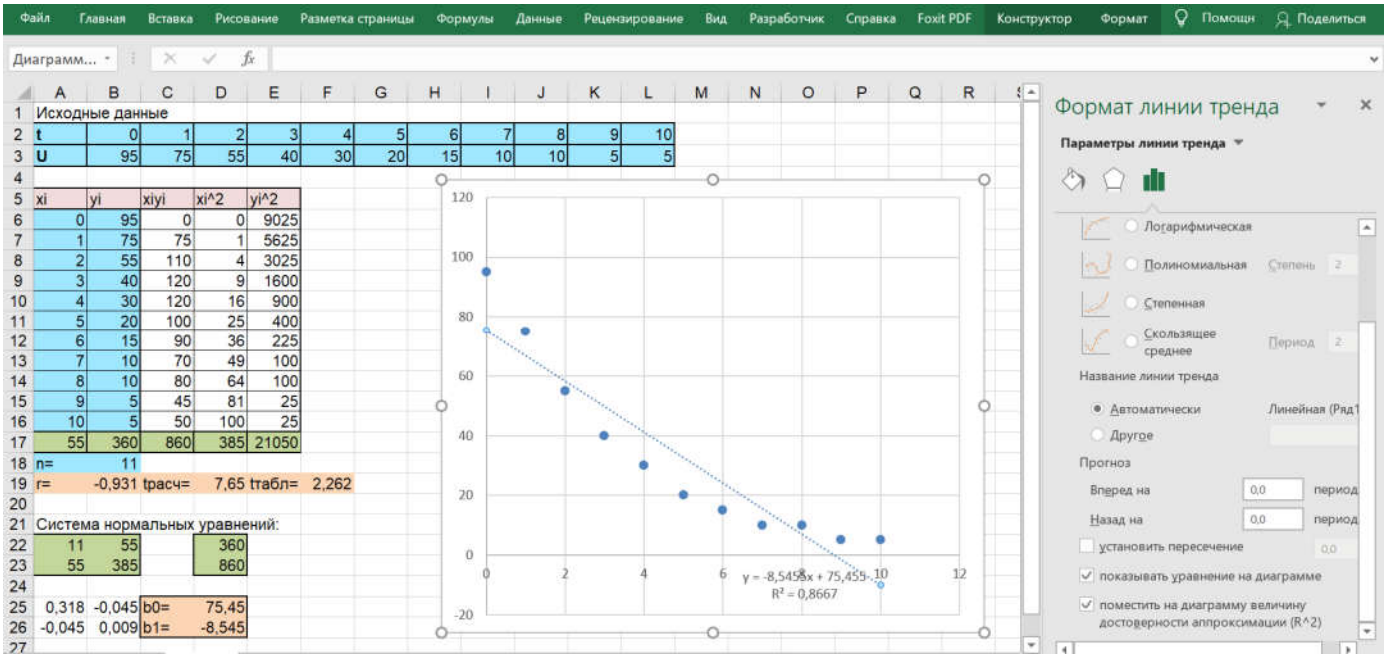
если $x = 10$, то $\hat{y} = -10,05.$



Согласно МНК, построенная прямая приближает экспериментальные данные наилучшим образом в том смысле, что будет наименьшей сумма квадратов отклонений от экспериментальных точек по вертикали.

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.

Ниже приведен фрагмент рабочего листа и даны рекомендации по выполнению пунктов 1-5 в Excel.



Методические указания по использованию EXCEL

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

2. 1) Функция КОРРЕЛ() вычисляет коэффициент корреляции между двумя массивами данных.

2) Функция СТЬЮДЕНТ.ОБР.2Х() вычисляет квантиль распределения Стьюдента. *Например,* в ячейке F19 использована формула

=СТЬЮДЕНТ.ОБР.2Х(0,05;B18-2).

4. В примере на фрагменте рабочего листа Excel система нормальных уравнений решена матричным методом: если A – матрица системы, X – столбец неизвестных, B – столбец свободных членов, то система равносильна матричному уравнению $AX = B$, а ее решение находится по формуле $X = A^{-1}B$.

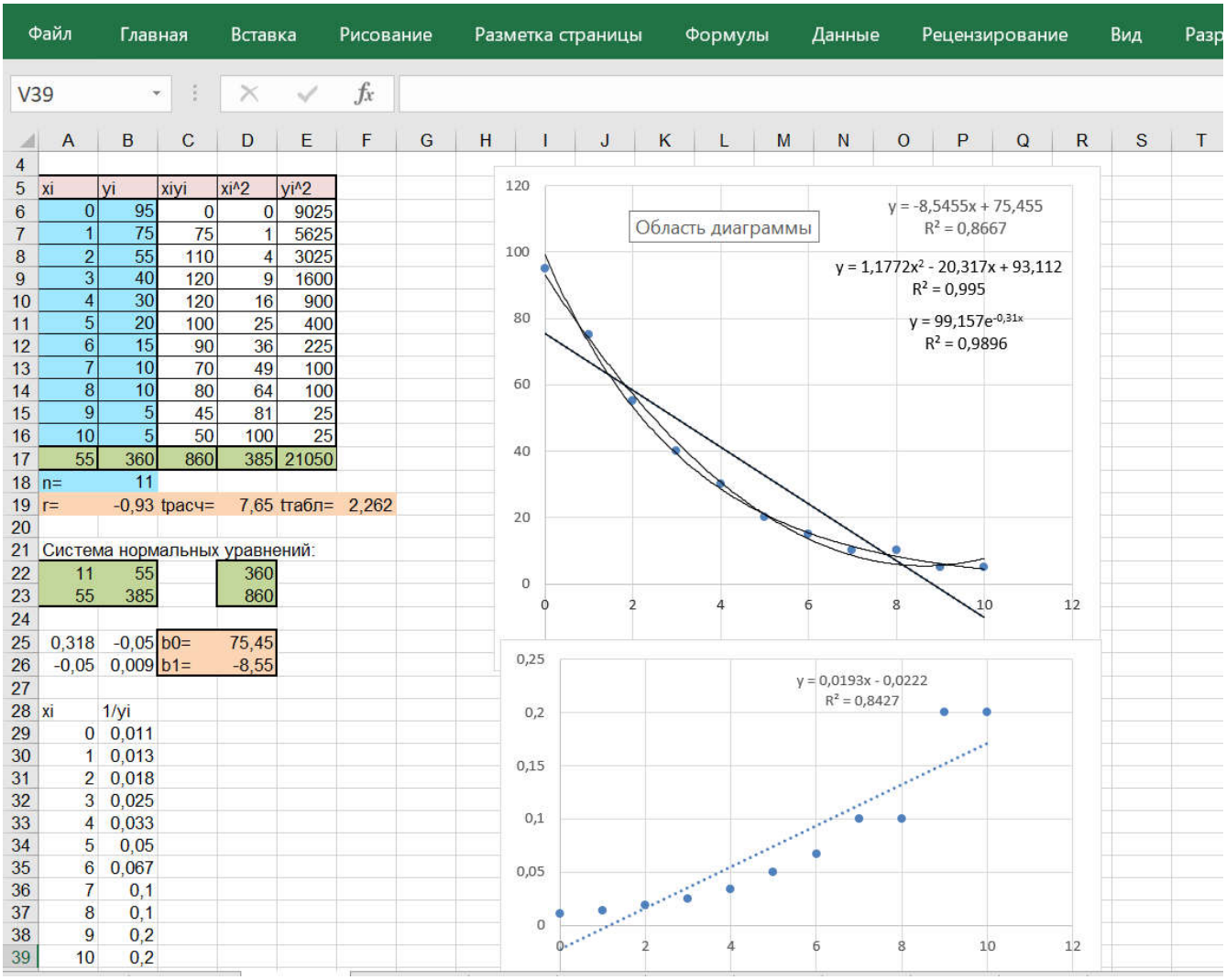
1) Функция МОБР() вычисляет обратную матрицу. *Например*, в примере для вычисления матрицы, обратной к матрице, записанной в массиве A22:B23, выделили массив A25:B26 для записи обратной матрицы, вызвали функцию =МОБР(A22:B23) и нажали сочетание клавиш **Ctrl+Shift+Enter** (три клавиши вместе!).

2) Функция МУМНОЖ() вычисляет произведение матриц. *Например*, в примере для вычисления произведения обратной матрицы на массив D22:D23 выделили массив D25:D26 для записи результата, вызвали функцию =МУМНОЖ(A25:B26;D22:D23) и нажали сочетание клавиш **Ctrl+Shift+Enter**.

6. С помощью Excel подберем наилучшую аппроксимирующую функцию для исходных данных.

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	$y = 99,157\,\mathrm{e}^{-0,31x}$	0,9896	
е) Степенная	–	–	есть значение $x = 0$
ж) Гиперболическая	$y = \dfrac{1}{0,0193x - 0,0222}$	0,8427	

Ниже приведен фрагмент рабочего листа и даны рекомендации по выполнению пункта 6 в Excel.



	A	B	C	D	E
4					
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
20					
21	Система нормальных уравнений:				
22	11	55		360	
23	55	385		860	
24					
25	0,318	-0,05	b0=	75,45	
26	-0,05	0,009	b1=	-8,55	
27					
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			

6. 1) Чтобы получить на диаграмме линейное 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)

2) Зависимости г), е) нельзя получить, если есть нулевые или отрицательные значения переменной *x*; зависимости д), е) нельзя получить, если есть нулевые или отрицательные значения переменной *y*. Если все значения переменной *y* отрицательны, для получения зависимостей д) и е) (с отрицательным коэффициентом) следует сделать замену $Y = |y|$.

3) Для получения гиперболической зависимости в) (если нет нулевых значений *x*) сделайте замену переменных $Y = y, X = \frac{1}{x}$, постройте новую диаграмму и добавьте на ней линейную линию тренда.

4) Для получения гиперболической зависимости ж) (если нет нулевых значений *y*) сделайте замену переменных $Y = \frac{1}{y}, X = x$, постройте новую диаграмму и добавьте на ней линейную линию тренда.

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.

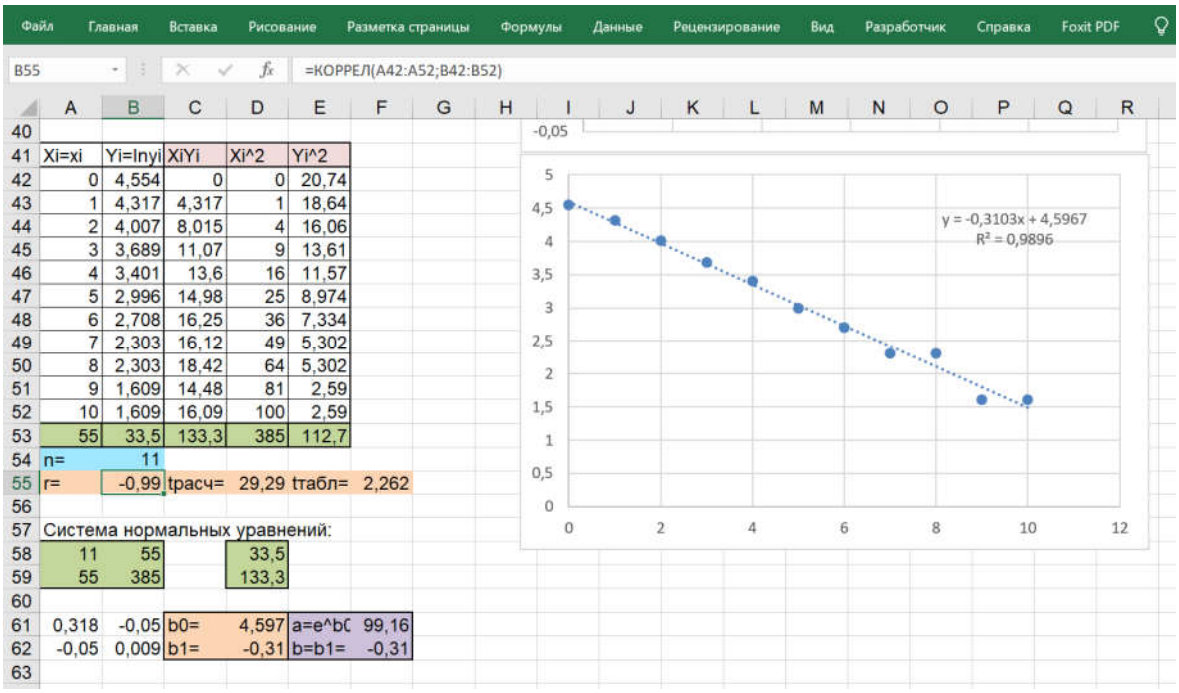
Таким образом, наилучшей аппроксимирующей функцией признаем экспоненциальную функцию $y = 99,157\,e^{-0,31x}$ с $R^2 = 0,9896$.

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

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

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 $a = e^{b_0}, b = b_1$.

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	F
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						

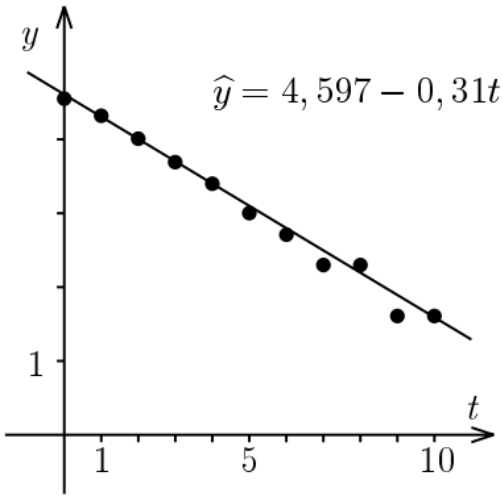
На диаграмме точки $(X_i; Y_i)$ располагаются вдоль прямой, коэффициент корреляции $r_{X;Y} = -0{,}99,$ а значит, вид зависимости y от x подобран правильно.

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

$$\begin{cases} nb_0 + b_1 \sum\limits_{i=1}^{n} X_i = \sum\limits_{i=1}^{n} Y_i, \\ b_0 \sum\limits_{i=1}^{n} X_i + b_1 \sum\limits_{i=1}^{n} X_i^2 = \sum\limits_{i=1}^{n} X_i Y_i; \end{cases} \qquad \begin{cases} 11b_0 + 55b_1 = 33{,}5, \\ 55b_0 + 385b_1 = 133{,}3. \end{cases}$$

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.