2. Статистический анализ данных в пакете STATISTICA

STATISTICA представляет собой интегрированную систему статистического анализа и обработки данных. Она состоит из следующих основных компонентов в рамках оной системы:

- электронных таблиц для данных (Spreadsheets) и специальных таблиц вывода численных результатов анализа;
- графической системы визуализации данных и результатов статистического анализа;
- набора специализированных статистических модулей, в которых собраны группы логически связанных между собой статистических методов.

Система включает в себя более тридцати специализированных статистических модулей и работает в мультизадачном режиме (одновременная работа с несколькими модулями). В ходе лабораторного практикума студент освоит работу следующих модулей:

- **Multiple Regression** (множественная регрессия);
- **Nonlinear Estimation** (нелиейное оценивание);
- **Basic Statistics / Tables** (базовые оценки статистики / таблицы).

2.1. Лабораторная работа №1: «Пакет STATISTICA 7»

Цель работы: познакомиться с интерфейсом пакета, освоить операции ввода/корректировки/сохранения таблиц данных и получить базовые навыки работы с модулями Nonlinear Estimation и Multiple Regression.

2.1.1. Общее знакомство с пакетом

Для запуска пакета необходимо выполнить двойной щелчок левой кнопкой мыши (ДЩ ЛКМ) в меню Программы ➔ Математика.

Структура окна (рис. 2.1) соответствует стандартам Windows приложений и включает следующие элементы:

- **строка заголовка (1)**: слева указано название пакета и имя открытого сейчас файла (по умолчанию имя - Spreadsheet1); справа размещены кнопки управления окном **(2):** «свернуть» в панель задач, «развернуть» на весь экран / «восстановить» исходный размер окна и «закрыть».
- **главное меню команд пакета (3)**:
    - **File** – операции с файлами (создать / сохранить / импортировать / экспортировать и т.д.)
    - **Edit** – редактирование таблиц и информации в них (оно подобно MS Office);

- **Insert** – вставка / копирование строк (Cases) и столбцов (Variables);
- **Statistics** – запуск встроенных статистических модулей;
- **Graphs** – построение графиков различного типа;
- **Data** – работа с таблицами данных (редактирование, расчет по формулам, стандартизация);
- **Window** – при нескольких запущенных модулях позволяет изменить способ расположения этих окон (каскад / по вертикали / по горизонтали), вывести на передний план конкретное или закрыть их все разом (на экране останется только таблица исходных данных).

- **панель инструментов (4)** в виде ряда кнопок, дублирующих функционал соответствующих пунктов меню команд (при задержании курсора на кнопке всплывает подсказка о выполняемой команде). По умолчанию включены стандартная (Standart Toolbar) и панель работы с таблицами данных (Spreadsheet Toolbar). Активные панели настраиваются через меню команд View ➔ Toolbars.
- **окно Data (5)**: в строке заголовка указывается имя открытой в данный момент таблицы, её размеры в скобках (10v by 10c – 10 столбцов на 10 строк). Строки нумеруются, а столбцы имеют названия (Var1, Var2 и т.д.). В пустую строку под заголовком таблицы можно вводить любую текстовую информацию как заметку, характеризующую данные в таблице.
- ниже располагается **панель анализа (Analysis Bar) (6)** с кнопкой стартового меню, дублирующего некоторые команды главного меню пакета. Рядом с кнопкой на эту же панель помещаются клавиши окон запущенных модулей, которые можно сделать активными щелчком мыши.
- **статусная панель (Status Bar) (7)** выводит информацию о состоянии процессов

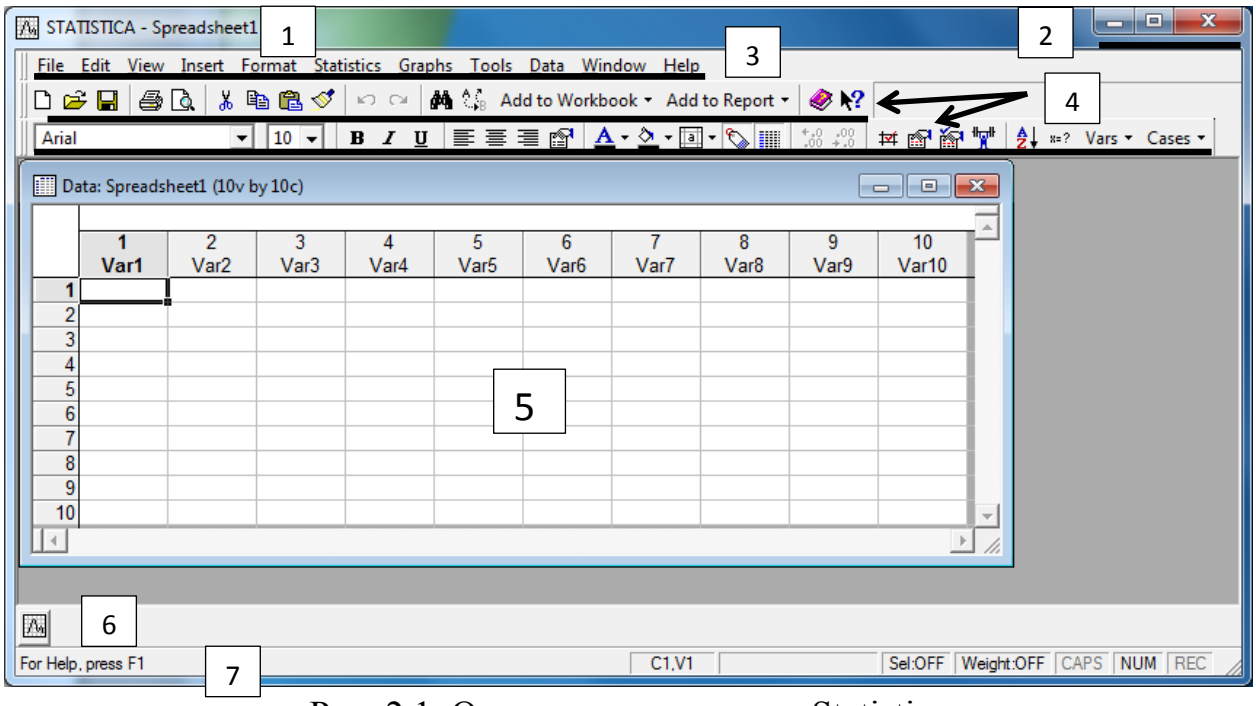

Рис. 2.1. Основное окна пакета Statistica

2.1.3. Работа с таблицами данных

Данные в пакете представляются в виде таблиц SpreadSheet и хранятся в памяти компьютера в специализированном формате с расширением файла .sta. **Столбцы** называются **переменными (Variables)**, а **строки – опытами / экспериментами (Cases)**. Команды операций с таблицами подробно рассмотрены и в приложении (**Операции с таблицами.htm**). Осваивать работу будем на конкретном примере. Предлагается простая задача поиска уравнения, чтобы сосредоточить внимание на инструментарии модулей пакета.

Задача. По экспериментальным значениям *константы скорости химической реакции* (K), полученным при различной *температуре* (t), *определить параметры уравнения Аррениуса* K_0 и E. Уравнение Аррениуса имеет следующий вид:

$$K = K_0 * e^{-E/RT}, \qquad (2.1)$$

где R – универсальная газовая постоянная [8.31 Дж/(моль*К)], T – температура процесса в К.

Поиск неизвестных коэффициентов в уравнении будем осуществлять с помощью двух модулей – Nonlinear Estimation (Нелинейная оценка) и Multiple Regression (Множественная линейная регрессия).

Подготовим исходную таблицу для дальнейших расчетов:

- Создаем таблицу для ввода данных содержащую 2 переменных и 11 строк.
- Под заголовком таблицы вводим сведения об авторе/-ах (ФИО, группа).
- Первой переменной присваиваем имя t (температура), второй – K (константа скорости).
- В качестве первого значения t вводим номер группы, остальные ячейки столбца заполняем по арифметической прогрессии с шагом 10.
- В качестве первого значения K вводим число, в 10 раз меньшее номера компьютера, за которым работаем (от 0,1 до 1,6). Остальные ячейки заполняем числами, близкими к геометрической прогрессии с шагом 2 – 3.
- Таблицу сохраняем в свой каталог с именем Arr№группы_№комп (к примеру, Arr220_02).

  Пример получаемой таблицы приведен ниже.

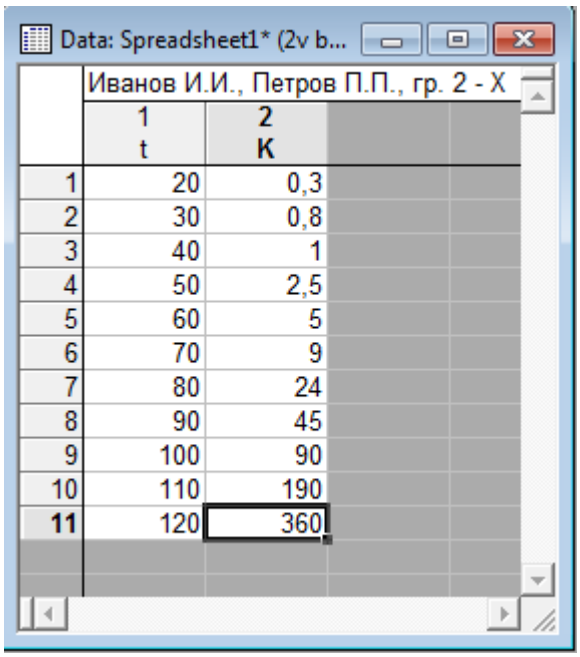

Рис. 2.2. Пример таблицы с экспериментальными данными

2.1.4. Визуализация исходных данных

Вполне логично перед поиском зависимости сначала построить график зависимости по исходным данным. Это позволит предугадать общий вид искомой зависимости. В STATISTICA графики можно строить как в виде точек, так и в виде линий. Построим график зависимости $K = f(t)$ в виде точек.

- Щелкаем ЛКМ в меню команд **Graph**, выбираем **Scatterplots**;
- В окне **2D Scatterplots** на закладку **Quick** щелкаем по кнопке **Variables**

- В Окне **Select Variables for Scatterplots** в левом списке (**X:**) выбираем переменную **t**, а правом (**Y:**) – **K**. Щелкаем ОК. Отключаем **Linear fit** и щелкаем по клавише окна **OK**.
- График появляется в окне Scatterplots, при этом на панели анализа появляется кнопка свернутого модуля 2D Scatterplots.

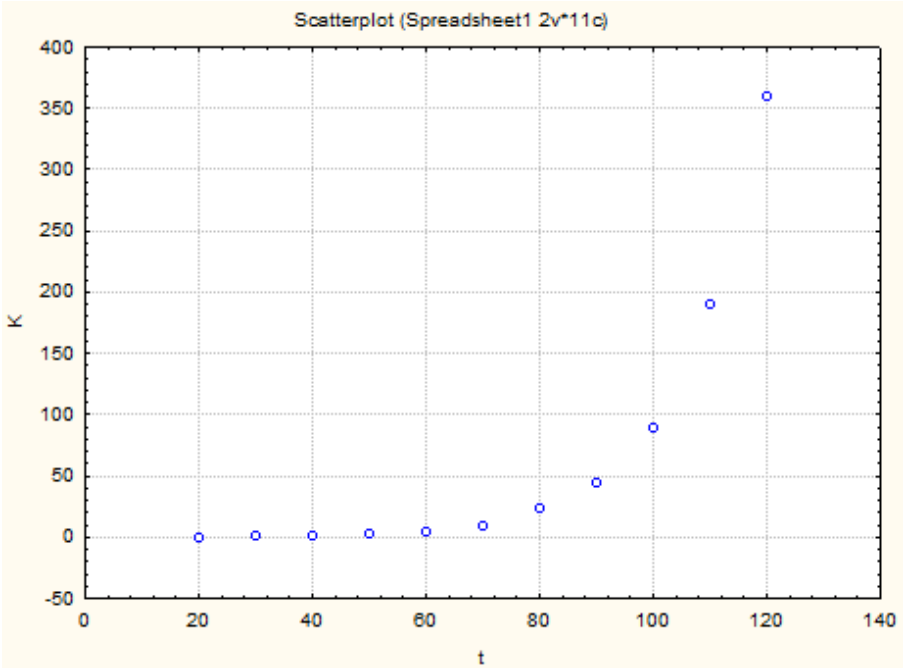

Рис. 2.3. Зависимость K=f(t)

- Мы видим, что график представляет собой нелинейную зависимость вида экспонента, что необходимо занести в отчет. Закрываем окно WorkBook1, не сохраняя его результаты.

2.1.5. Модуль нелинейного оценивания (Nonlinear Estimation)

Краткий математический анонс. Одна из наиболее распространенных задач статистического исследования состоит в изучении связи между некоторыми наблюдаемыми переменными. Переменные, варьируемые в эксперименте, называют факторами. Переменная, значение которой измеряют – откликом. При обработке экспериментальных данных часто предполагается нелинейная зависимость между откликом Y и одним / несколькими факторами (X_1, X_2…X_m), причем вид зависимости $Y = f(X_1, X_2…X_m)$ может быть известен и по экспериментальным данным необходимо только оценить неизвестные параметры данной зависимости. *Например, в рассмотренной выше задаче известно, что константа скорости химической реакции имеет вид экспоненциального уравнения (2.1), что мы подтвердили*

графически, и нам требуется по экспериментальным данным оценить K_0 и E.

Для решения этой задачи используются численные методы оптимизации, относящиеся к классу методов нелинейного программирования: квазиньютоновский метод, симплексный метод, метод Розенблока, метод Хука-Джива и комбинации этих методов. В качестве минимизируемой функции часто, как и в аппроксимации, используется сумма квадратов остатков.

Поиск нелинейной зависимости.

Решим подзадачу №1 – найдем оценки параметров K_0 и E. Исходные данные для расчета – подготовленная чуть выше таблица Arr№группы_№комп.sta. Отклик – K, а фактор – t (°C).

1. ***Вызов стартовой панели модуля***
   - Прежде всего, в пакете должна быть открыта таблица с исходными обрабатываемыми данными.
   - Щелкаем меню команд **Statictics**, выбираем **Advanced Linear / Nonlinear Estimation**, в подменю щелкаем **Nonlinear Estimation**.
   - В окно стартовой панели модуля выбираем **User-specified regression & custom loss function** и щелкаем **ОК**.
   - В окне щелкаем кнопку $Function\ to\ be\ estimated\ \&\ loss\ function$, в поле **Estimated Function** вводим функцию в компьютерном виде:

$$\boldsymbol{K = K0 * \exp(-E/(8,31 * (t + 273)))}$$

2. ***Выбор минимизируемой функции и численного метода поиска оценок:***
   - В поле **Loss function** не изменяем функцию, т.к. по умолчанию программа производит поиск оценок параметров из условия минимума суммы квадратов отклонений экспериментальных (OBS) и расчетных (PRED) значений, что нас вполне устраивает. Щелкаем **ОК**.

     Примечание: в нашем случае зависимость представляет собой грубо геометрическую прогрессию. Поиск неизвестных коэффициентов уравнения такой зависимости следует проводить не из разницы абсолютных, а из разницы относительных величин, т.е. в поле ***Loss function*** *минимизируемая функция должна выглядеть как ((OBS-PRED)/OBS)**2.*

- В окне **User-specified regression** щелкаем **OK**. Появляется окно **Model Estimation**: в информационной части приводятся искомая функция (Model), число оцениваемых параметров (Number of parameter to be estimate), зависимая (Depended) и независимая (Independent) переменная, число опытов (Number of Valid Cases).
- Ниже выбираем вкладку **Advanced**, в поле **Estimation Method** выбираем численный метод Хука-Дживса (**Hoove-Jeevs pattern moves**).
  Примечание: *Число итераций **Maxium Number of iterations** и точность метода **Convergence** не меняем. В данном окне можно также задать начальные значения (кнопка **Start Values**) и начальный шаг поиска (кнопка **Initial Step Values**) для оцениваемых параметров в виде чисел с плавающей запятой (мантисс, к примеру, 5.2E9, что означает $5.2*10^9$).* Однако в данной задаче в силу простоты искомой функции мы их не задаем.
  Щелкаем **OK** в окне **Model Estimation**.

3. Поиск оценок параметров нелинейной функции

- В окне **Parameter Estimation** отображается процесс поиска параметров. На каждой итерации выводятся: номер шага, значения минимизируемой функции (Loss) и значения искомых параметров (Parameters).
- Если заданного числа итераций окажется недостаточно на нахождение параметров с заданной точностью, программа запросит еще 30 итераций. В нашей задаче мы соглашаемся с указанным запросом до тех пор, пока будет не получено окно результатов **Results.**
- В информационной части этого окна приводится информация о финальном значении минимизирующей функции (Final Values) и коэффициента детерминации (R).

4. Вывод и анализ полученных результатов

- Для вывода найденных оценок параметров следует щелкнуть кнопку **Summary: Parameter Estimates**. В окне будут приведены оценки параметров K0 и E, а также информация о точности подгонки нелинейной функции к экспериментальным данным.

	Model: K=K0*exp(-E/(8.31*(t+273))) (Spreadsheet1.sta) Dep. var: K Loss: (OBS-PRED)**2 Final loss: 34,845931044 R=,99986 Variance explained: 99,973%	
N=11	K0	E
Estimate	2,785835E+13	81876,71

Рис. 2.4. Окно результатов поиска

Фактически, остается только перенести эти коэффициенты в уравнение Аррениуса, осуществив их перевод в алгебраическую форму:

Примечание:

- *Более наглядно о качестве подгонки можно судить по расположению экспериментальных точек относительно графика функций с найденными параметрами. Для вывода графика следует вернуться в окно* ***Results*** *(развернув его с панели анализа) и щелкать кнопку* ***Fitted 2D function & observed values****. Появится окно с графиком вида:*

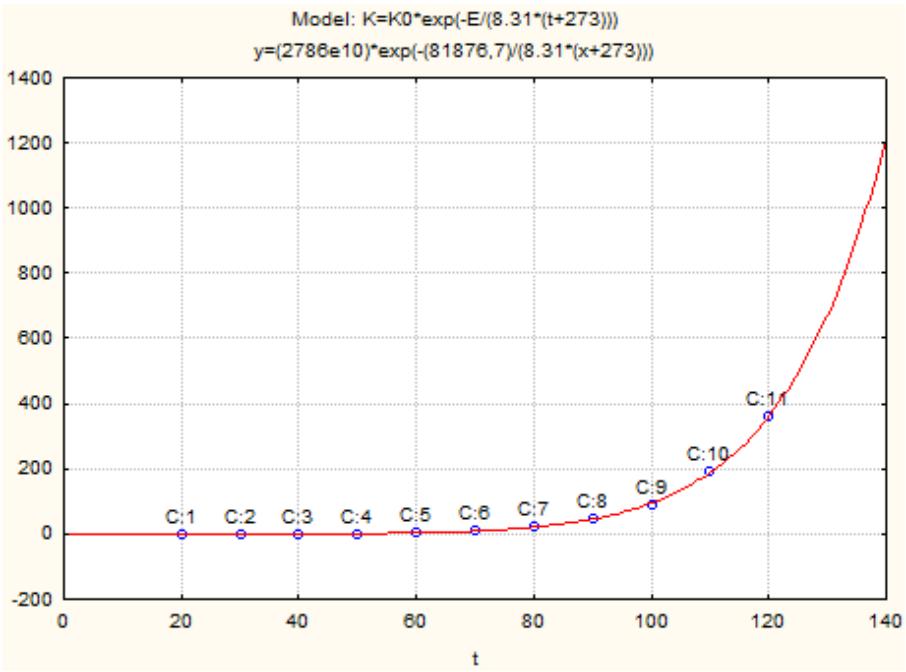

Рис. 2.5. График визуальной оценки качества подгонки

где точки с определенным номером соответствуют исходным данным в соответствующей строке, а линия визуализирует расчетную функцию. Над графиком приведено рассчитанное уравнение в общем виде и с найденными оценками.

- *Для количественного сравнения расчетных и экспериментальных значений в окне* ***Results*** *следует щелкнуть кнопку* ***Observed, predicted, residual values****, что откроет окно с таблицей. В колонке Observed*

представлены исходные экспериментальные данные, Predicted – расчетные, а Residuals – разница между двумя указанными колонками.

Model is: K=K0*exp(-E/(8.31*(t+273)))
Dep. Var. : K

	Observed	Predicted	Residuals
1	0,3000	0,0693	0,23069
2	0,8000	0,2103	0,58972
3	1,0000	0,5943	0,40573
4	2,5000	1,5748	0,92516
5	5,0000	3,9361	1,06391
6	9,0000	9,3260	-0,32604
7	24,0000	21,0428	2,95718
8	45,0000	45,3983	-0,39826
9	90,0000	93,9873	-3,98729
10	190,0000	187,3252	2,67481
11	360,0000	360,4791	-0,47913

Рис. 2.6. Количественная оценка точности подгонки

2.1.6. Нахождение оценок с помощью линейного регрессионного анализа

1. Краткий математический анонс

Регрессионный анализ состоит в установлении (идентификации) функциональной зависимости между откликом Y и одним / несколькими факторами (X_1, X_2, … X_m). В линейном регрессионном анализе эта зависимость предполагается линейной. В самом простом случае имеются де переменные X и Y. Требуется по m парам наблюдений (X_1, Y_1), (X_2, Y_2)…(X_m, Y_m) подобрать прямую линию, которая «наилучшим образом» приближает наблюдаемые значения. Как правило, линия подбирается из условия минимума суммы квадратов отклонений расчетных значений отклика от экспериментальных значений по всем опытам, т.е. методом наименьших квадратов (МНК). Математически задача регрессионного анализа может быть сформулирована следующим образом. Значениям независимой переменной X отвечают значения зависимой переменной Y (регрессия):

$$Y_i = \beta_0 + \beta_1 * X_i + \varepsilon_i,\ i = 1 \dots m, \quad (2.2)$$

где ε_i – независимые случайные ошибки со средним 0, которые интерпретируются как ошибки наблюдений; β_0, β_1 – неизвестные параметры, описывающие прямую линию, которые следует определить по наблюдениям (X_i, Y_i), i = 1…m. По результатам наблюдений можно получить лишь приближенные значения (оценки) параметров β_0 и β_1, обозначаемые b_0 и b_1. Уравнение связи, в которые входят данные оценки параметров, называют приближенной (выборочной) регрессией:

$$\hat{Y} = b_0 + b_1 * X, \quad (2.3)$$

где коэффициенты b_0 и b_1 рассчитываются из условия:

$$\Phi = \sum_{i=1}^{m} (\widehat{Y}_i - Y_i)^2 \quad (2.4)$$

Разность $\widehat{Y}_i - Y_i$ называют остатком i-го опыта. По его величине можно судить о качестве подгонки линейно зависимости. Выборочная регрессия (2.3) позволяет найти значение отклика при любом факторе, не прибегая к выполнению эксперимента.

Подзадача №2. Путем логарифмирования зависимость (2.1) приводим к линейному виду:

$$\ln(K) = \ln(K_0) - \frac{E}{R} * \frac{1}{T} \text{ или в виде регрессии } Y = b_0 + b_1 * X \quad (2.5)$$

В таком случае для решения задачи необходимо найти значения коэффициентов линейной регрессии b_0 и b_1 и от них вернуться к исходным параметрам:

$$K_0 = e^{b_0} \quad (2.6)$$

$$E = b_1 * R = 8.31 * b_1 \quad (2.7)$$

Обработка будет выполняться в модуле множественная регрессия (Multiple Regression).

2. Подготовка исходных данных. В качестве экспериментальных данных будут использованы данные из файла, подготовленного по п.2.1.3.

- Добавим две новые переменные X и Y (можно через меню Data, через кнопку Vars на панели инструментов или просто кликнув дважды по любому серому полю вне таблицы с данными);
- Вычислим их по формулам: $Y = log(K)$, $X = 1/(273 + t)$. Данные формулы вводятся в **спецификациях** соответствующих **переменных** в поле **Long Name (or formula)** в виде =*некое выражение*.
- Сохраним изменения в исходный файл.

Data: Spreadsheet1.sta* (4v by 11c)

	Иванов И.И., Петров П.П., гр. 2 - X			
	1 t	2 K	3 X	4 Y
1	20	0,3	0,003413	-1,20397
2	30	0,8	0,0033	-0,22314
3	40	1	0,003195	0
4	50	2,5	0,003096	0,916291
5	60	5	0,003003	1,609438
6	70	9	0,002915	2,197225
7	80	24	0,002833	3,178054
8	90	45	0,002755	3,806662
9	100	90	0,002681	4,49981
10	110	190	0,002611	5,247024
11	120	360	0,002545	5,886104

Рис. 2.7. Пример исходных данных для линейного регрессионного анализа

3. Построение графика Y=f(X) в виде прямой:

- Щелкаем ЛКМ в меню команд **Graph**, выбираем **Scatterplots** (или разворачиваем модуль из панели анализа по кнопке **2D Scatterplots**);
- В окне **2D Scatterplots** кликаем кнопку **Variables**;
- В Окне **Select Variables for Scatterplots** в левом списке (**X:**) выбираем переменную **X**, а правом (**Y:**) – **Y**. Щелкаем ОК.
- Включаем **Linear fit** и щелкаем по клавише окна **OK**.

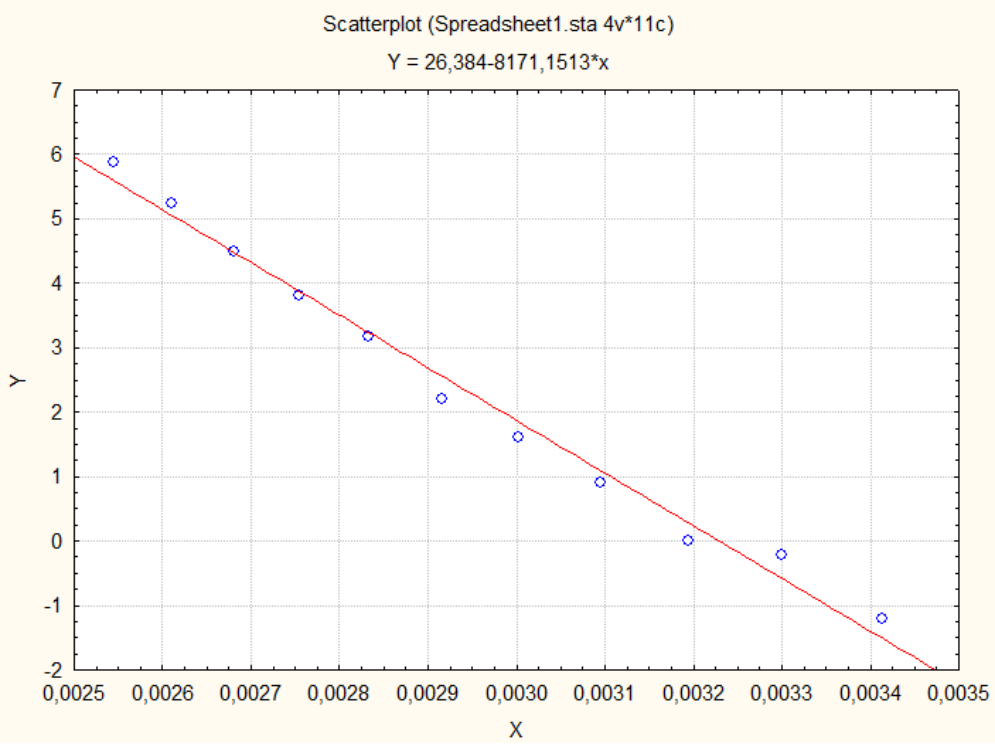

Рис. 2.8. Зависимость Y = f(X)

Примечание:

- *Для построения графика с указанием доверительных границ во вкладке* ***Advanced*** *в группе* ***Regression bands*** *включается опция* ***Confidence Level****.*

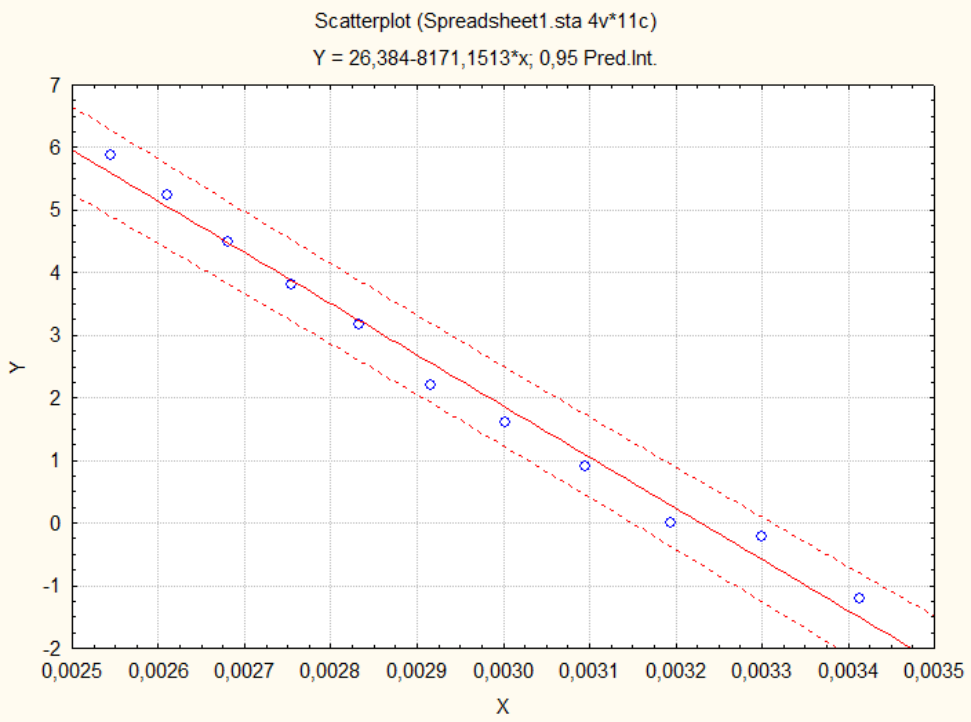

Рис. 2.9. Зависимость Y=f(X) с доверительными границами ($p = 0.95$)

4. Определение параметров линейной выборочной регрессии в модуле Multiple Regression

- В меню команд выбираем **Statistics**, далее – **Multiple Regression**
- В окне **Multiple Linear Regression** на закладке **Quick** щелкаем по кнопке **Variables**, в левом списке открывшегося окна ЛКМ выбираем зависимую (**Depended**) переменную – **Y**, а в правом – независимую (**Independed**) – **X**, щелкаем по кнопке **OK**. Проверяем выбор в окне **Multiple Linear Regression** и щелкаем **OK**.
- Для вывода результатов в **Multiple Linear Regression** щелкаем по кнопке **Summary: Regression Results**:

Regression Summary for Dependent Variable: Y (Spreadsheet1.sta)
R= ,99416495 R?= ,98836394 Adjusted R?= ,98707105
F(1,9)=764,46 p<,00000 Std.Error of estimate: ,26842

N=11	Beta	Std.Err. of Beta	B	Std.Err. of B	t(9)	p-level
Intercept			26,38	0,8728	30,2288	0,000000
X	-0,994165	0,035957	-8171,15	295,5333	-27,6488	0,000000

Рис. 2.10. Результаты поиска в линейном виде

Коэффициенты регрессии (2.3) содержатся в столбце B: b_0 – в строке Intercept, а b_1 – в строке X. Т.о., регрессия принимает вид:

$$\hat{Y} = 26.38 - 8171.15 * X$$

Примечание:

***Для количественной оценки**, насколько точно данным уравнением описываются экспериментальные данные, используются коэффициенты корреляции (R) и **детерминации (R?)**, которые должны быть как можно ближе к 1. **Для визуальной оценки** соответствия наблюдаемых (Observed) Y и предсказанных (Predicted) $\hat{Y}$ значений отклика можно вывести график. Для этого*

- *щелкаем на панели анализа по кнопке **Multiple Regression**;*
- *в окне **Multiple Linear Regression Results** кликаем **OK**;*
- *в окне **Residual Analysis** выбираем вкладку **Scatterplots** и щелкаем по кнопке **Predicted vs Observed***

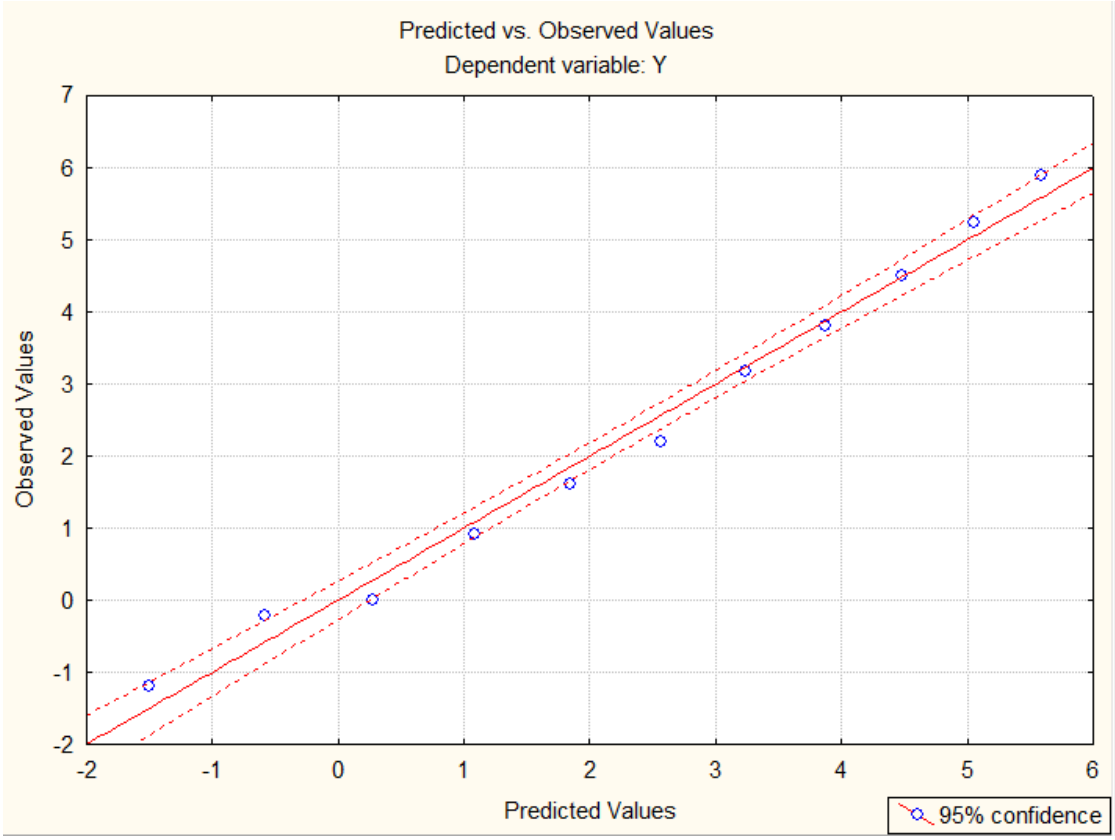

Рис. 2.11. График визуальной оценки качества подгонки линейной регрессией

Если все точки находятся внутри доверительной трубки, найденная регрессия с вероятностью 95% точно отражает экспериментальные данные. Если точка лежит прямо на линии, то это означает, что при одном и том же факторе экспериментальное значение отклика совпадает с расчетным.

5. Параметры уравнения Аррениуса

- В исходной таблице создадим еще 2-е переменные – K0 и E;
- Для расчета используются формулы (2.6) и (2.7):

- $K0 = \exp(26.38) = 2.86E + 11$
- $E = 8.31 * 8171.15 = 69902.26$ Дж/моль

Data: Spreadsheet1.sta* (6v by 11c)

	Иванов И.И., Петров П.П., гр. 2 - X					
	1 t	2 K	3 X	4 Y	5 K0	6 E
1	20	0,3	0,003413	-1,20397	2,8621E+11	67902,26
2	30	0,8	0,0033	-0,22314	2,8621E+11	67902,26
3	40	1	0,003195	0	2,8621E+11	67902,26
4	50	2,5	0,003096	0,916291	2,8621E+11	67902,26
5	60	5	0,003003	1,609438	2,8621E+11	67902,26
6	70	9	0,002915	2,197225	2,8621E+11	67902,26
7	80	24	0,002833	3,178054	2,8621E+11	67902,26
8	90	45	0,002755	3,806662	2,8621E+11	67902,26
9	100	90	0,002681	4,49981	2,8621E+11	67902,26
10	110	190	0,002611	5,247024	2,8621E+11	67902,26
11	120	360	0,002545	5,886104	2,8621E+11	67902,26

Т.о., уравнение Аррениуса с найденными параметрами для рассматриваемого примера записывается в виде:

$$K = 2.86 * 10^{11} * e^{-69902.26/RT}$$

2.1.7. Требования к оформлению лабораторной работы

Отчет по лабораторной работе должен содержать:

- название работы;
- цель работы;
- формулировку подзадач №1 и №2
- экспериментальные данные (достаточно представить таблицу всего один раз, т.е. или в п. 2.1.3, или в п.2.1.6.2)
- схему выполнения каждого действия в пакете (например, создали таблицу из 2-х столбцов и 11 строк: меню File ➔ New ➔ Spreadsheet ➔ NoV – 2, NoC – 11 ➔ OK); если действие в работе выполняется не единожды, достаточно подробно описать его всего 1 раз.
- схему решения подзадачи №1 в модуле Nonlinear Estimation с полученными результатами
- схему решения подзадачи №2 в модуле Multiple Regression с полученными результатами
- вывод, в котором представлены результаты по обеим подзадачам.