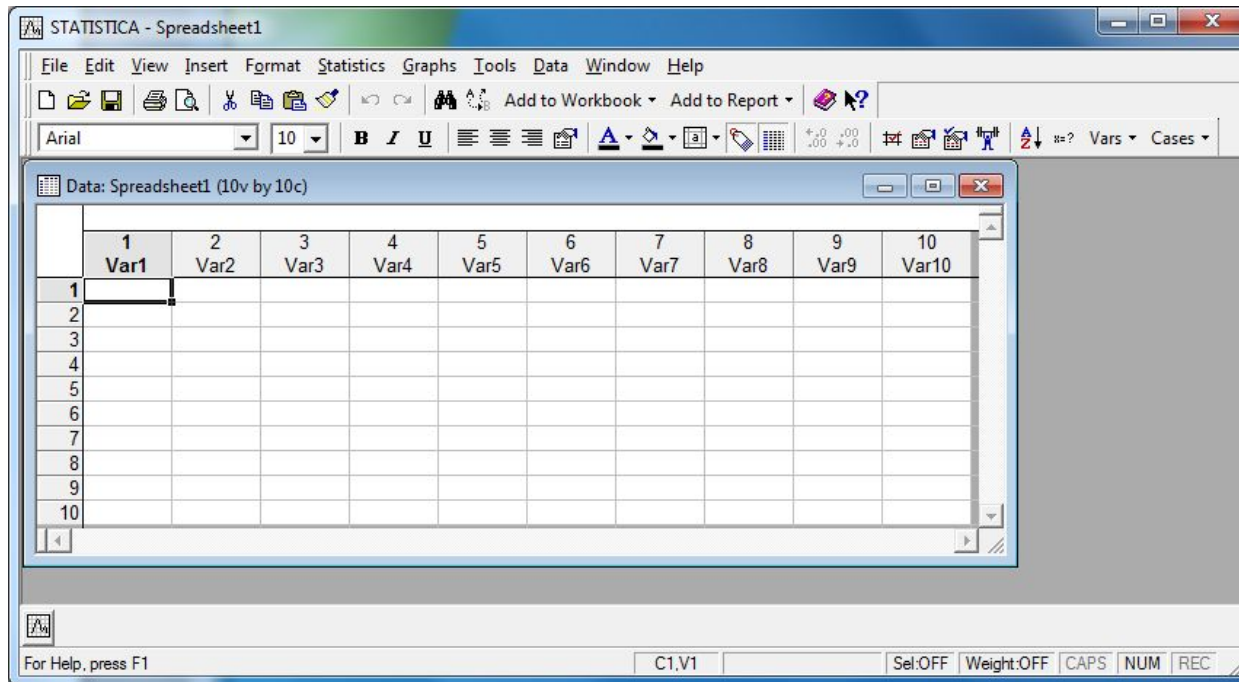


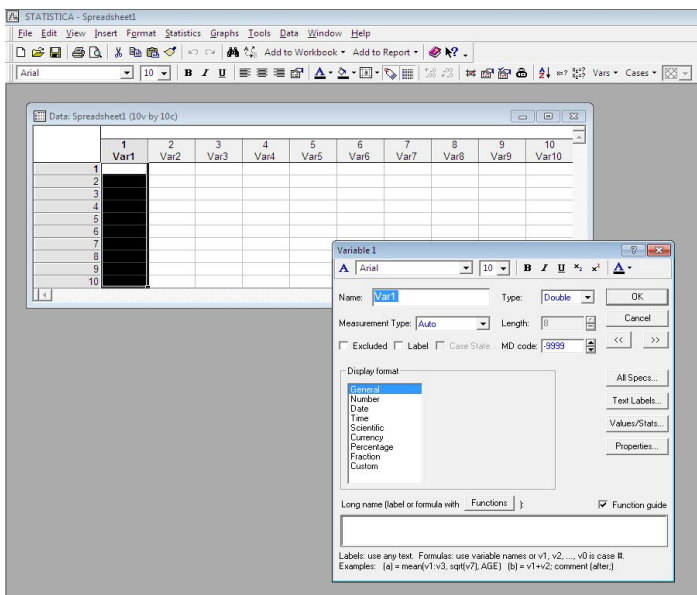
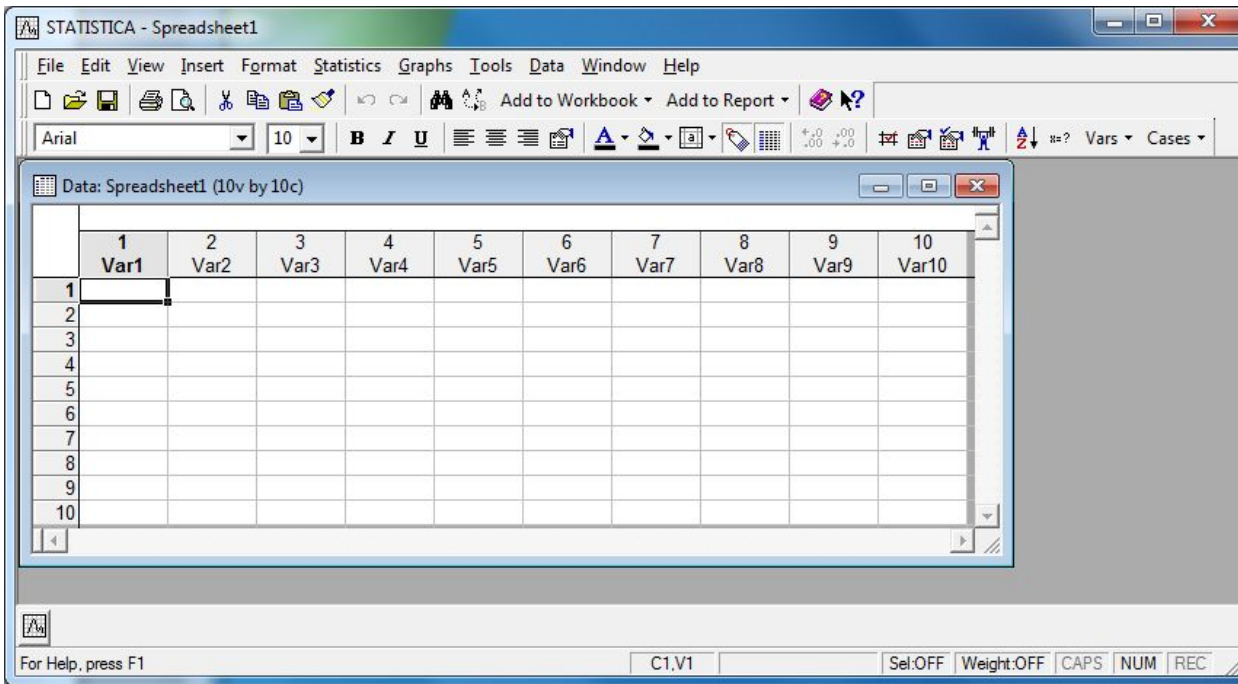
Статистический анализ данных в пакете STATISTICA

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:

$$K = K_0 * e^{-E/RT} \quad (1)$$

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





Data: Spreadsheet1* (2v b...)

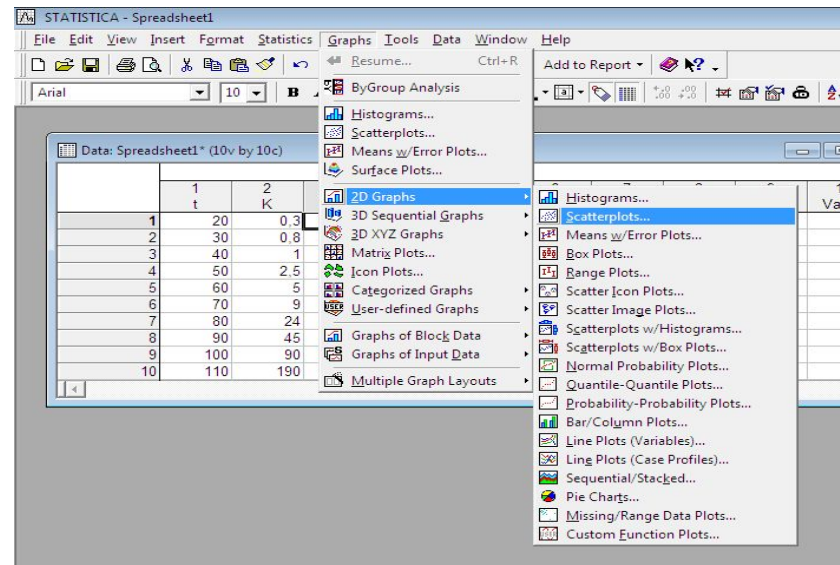
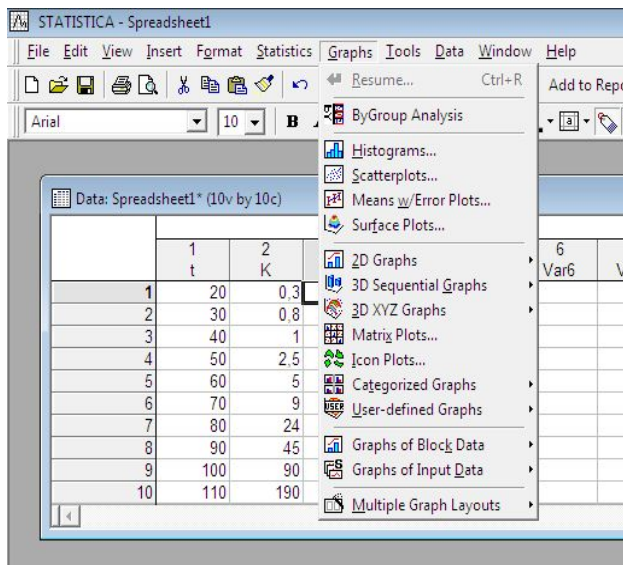
	Иванов И.И., Петров П.П., гр. 2 - X	
	1 t	2 K
1	20	0,3
2	30	0,8
3	40	1
4	50	2,5
5	60	5
6	70	9
7	80	24
8	90	45
9	100	90
10	110	190
11	120	360

Рис. 1. Пример таблицы с экспериментальными данными

Визуализация исходных данных

Вполне логично перед поиском зависимости сначала построить график зависимости по исходным данным. Это позволит предугадать общий вид искомой зависимости.

- 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 **ОК**.



STATISTICA - Spreadsheet

File Edit View Insert Format Statistics Graphs Tools Data Window Help

Arial 10 B I U

Data: Spreadsheet* (10v by 10c)

	1	2	3	4	5	6	7	8	9	10
	t	K	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10
1	20	0.3								
2	30	0.8								
3	40	1								
4	50	2.5								
5	60	5								
6	70	9								
7	80	24								
8	90	45								
9	100	90								
10	110	190								

2D Scatterplots

Quick | Advanced | Appearance | Categorized | Options 1 | Options 2

Variables:

X: none
Y: none

Graph type:

Regular (selected)
Multiple

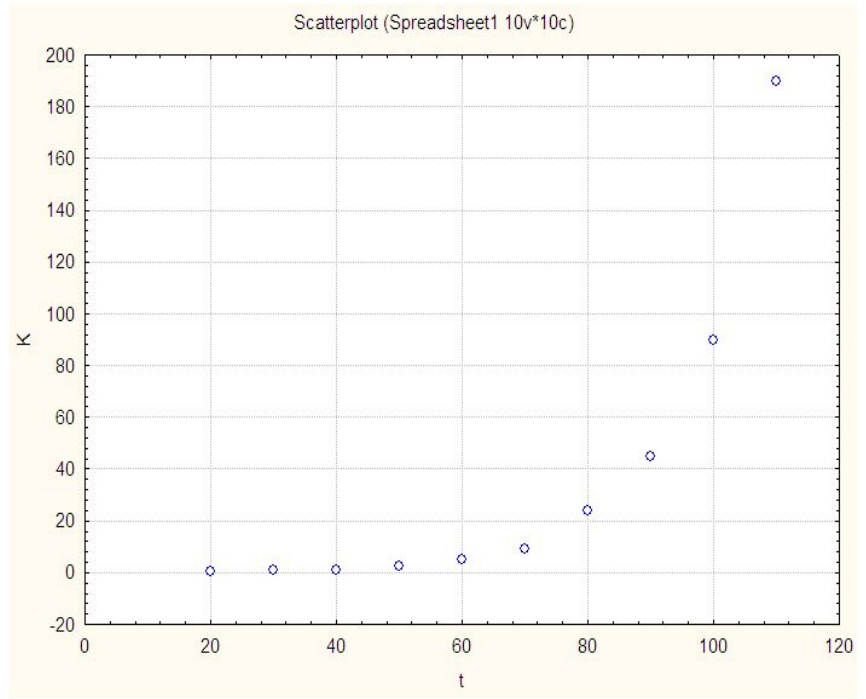
Regression bands:

Off level: 0.1
 Confidence 95
 Prediction

Fit type:

Linear

OK Отмена



STATISTICA - Spreadsheet

File Edit View Insert Format Statistics Graphs Tools Data Window Help

Arial 10 B I U

Data: Spreadsheet* (10v by 10c)

	1	2	3	4	5	6	7	8	9	10
	t	K	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10
1	20	0.3								
2	30	0.8								
3	40	1								
4	50	2.5								
5	60	5								
6	70	9								
7	80	24								
8	90	45								
9	100	90								
10	110	190								

2D Scatterplots

Quick | Advanced | Appearance | Categorized | Options 1 | Options 2

Variables:

Select Variables for Scatterplot

1:t
2:K
3:Var3
4:Var4
5:Var5
6:Var6
7:Var7
8:Var8
9:Var9
10:Var10

1:t
2:K

X: 1
Y: 2

Show appropriate variables only

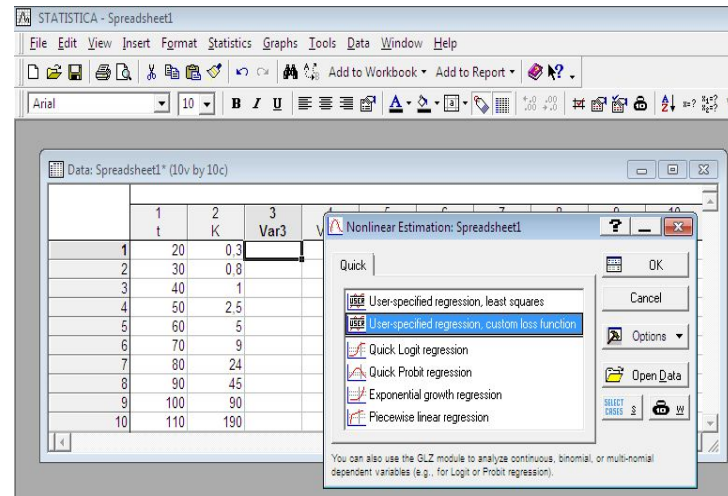
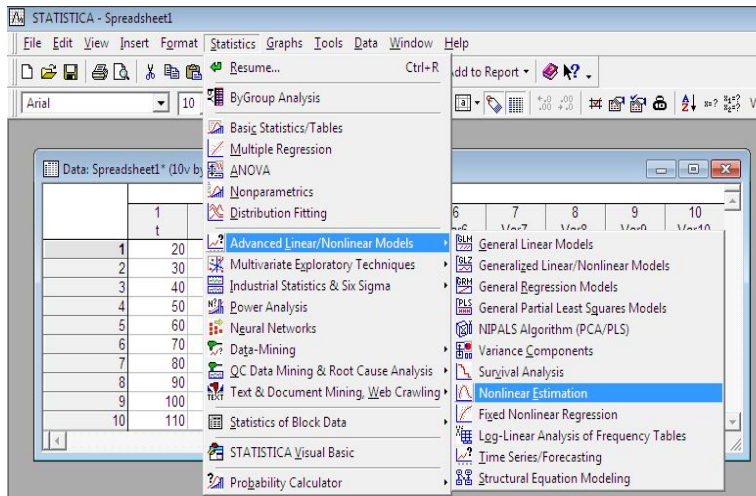
OK Отмена

Модуль нелинейного оценивания

Решим задачу первым способом – найдем оценки параметров K_0 и E с помощью методов оптимизации.

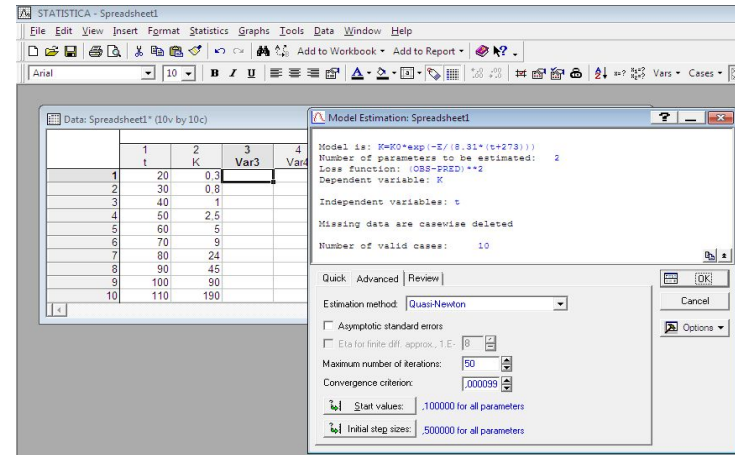
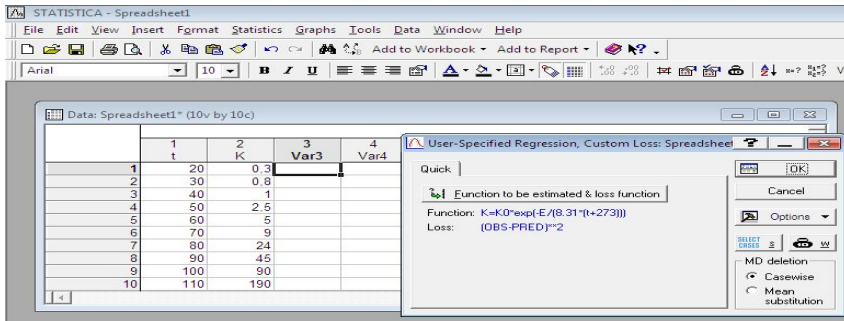
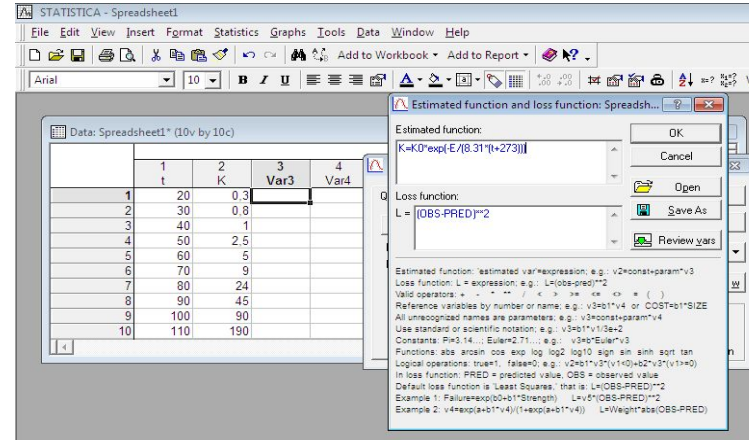
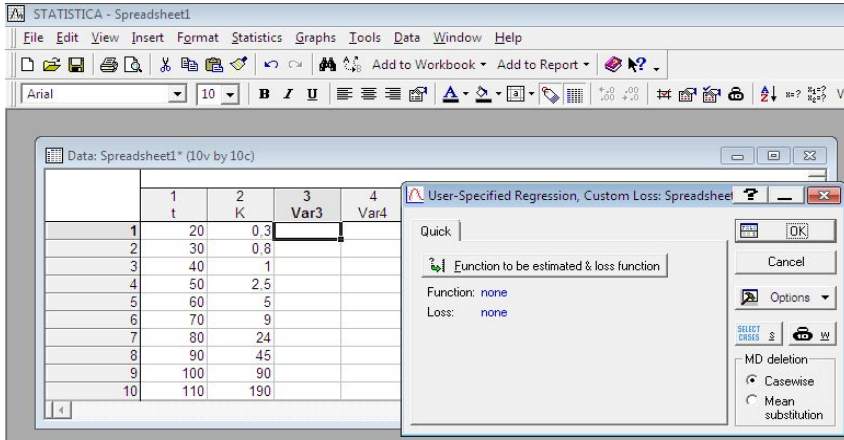
Вызов стартовой панели модуля

- Прежде всего, в пакете должна быть открыта таблица с исходными обрабатываемыми данными.
- Щелкаем меню команд **Statistics**, выбираем **Advanced Linear / Nonlinear Estimation**, в под меню щелкаем **Nonlinear Estimation**.
- В окне стартовой панели модуля выбираем **User-specified regression & custom loss function** и щелкаем **OK**.



- В окне щелкаем кнопку *Function to be estimated & loss function*, в поле **Estimated Function** вводим функцию в компьютерном виде:

$$K = K0 * \exp(-E / (8.31 * (t + 273)))$$



Выбор минимизируемой функции и численного метода поиска оценок:

- В окне **User-specified regression** щелкаем **ОК**. 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**). 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 $5.2 \cdot 10^9$). Однако в данной задаче в силу простоты искомой функции мы их не задаем. Щелкаем **ОК** в окне **Model Estimation**.

Нахождение оценок с помощью линейного регрессионного анализа

Регрессионный анализ состоит в установлении (идентификации) функциональной зависимости между откликом Y и одним / несколькими факторами (X_1, X_2, \dots, X_m). В линейном регрессионном анализе эта зависимость предполагается линейной. В самом простом случае имеются две переменные X и Y . Требуется по m парам наблюдений $(X_1, Y_1), (X_2, Y_2) \dots (X_m, Y_m)$ 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):

$$Y_i = \beta_0 + \beta_1 * X_i + \varepsilon_i, \quad i = 1 \dots m, \quad (2)$$

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 $(X_i, Y_i), i = 1 \dots m$. 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:

$$\hat{Y} = b_0 + b_1 * X, \quad (3)$$

где коэффициенты b_0 и b_1 рассчитываются из условия:

$$\Phi = \sum_{i=1}^m (\hat{Y}_i - Y_i)^2 \quad (4)$$

Разность $\hat{Y}_i - Y_i$ называют остатком i -го опыта. По его величине можно судить о качестве подгонки линейно зависимости. Выборочная регрессия (3) позволяет найти значение отклика при любом факторе, не прибегая к выполнению эксперимента.

Решим задачу вторым способом. Путем логарифмирования зависимость (1) приводим к линейному виду:

$$\ln(K) = \ln(K_0) - \frac{E}{R} * \frac{1}{T} \quad \text{или в виде регрессии } Y = b_0 + b_1 * X \quad (5)$$

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

$$K_0 = e^{b_0} \quad (6)$$

$$E = b_1 * R = 8.31 * b_1 \quad (7)$$

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

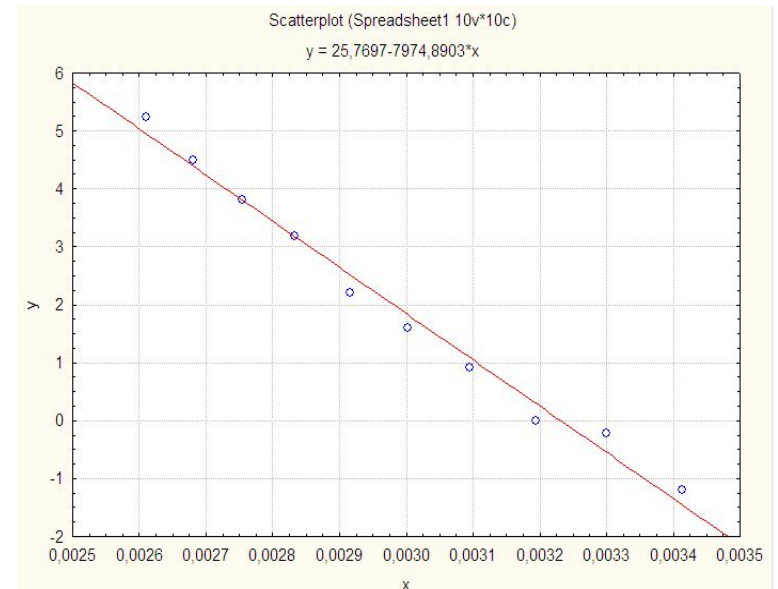
Построение графика $Y=f(X)$ в виде прямой:

- 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 **OK**.
- Включаем **Linear fit** и щелкаем по клавише окна **OK**.

The screenshot shows the STATISTICA software interface. The main window displays a data table with 10 rows and 6 columns. The columns are labeled 1-t, 2-K, 3-x, 4-y, 5-Var5, and 6-Var6. The data values are as follows:

	1-t	2-K	3-x	4-y	5-Var5	6-Var6
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		

The **2D Scatterplots** dialog box is open, showing the **Variables** section with **X:** none and **Y:** none. The **Graph type** section has **Regular** selected. The **Regression bands** section has **Off** selected. The **Select Variables for Scatterplot** dialog box is also open, showing the selection of **X:** 3 and **Y:** 4.



- 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**:

The screenshot displays the STATISTICA software interface. The main window shows a spreadsheet with the following data:

	1	2	3	4	5	6	7	8	9	10
	t	K	x	y	Var5	Var6	Var7	Var8	Var9	Var10
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						

Two dialog boxes are open over the spreadsheet:

- Select dependent and independent variable lists:** This dialog shows a list of variables on the left (1-t, 2-K, 3-x, 4-y, 5-Var5, 6-Var6, 7-Var7, 8-Var8, 9-Var9, 10-Var10). Variable 4-y is selected in the dependent list, and variable 3-x is selected in the independent list. The 'Dependent var. (or list for batch):' field contains '4' and the 'Independent variable list:' field contains '3'.
- Multiple Linear Regression: Spreadsheet1:** This dialog is on the 'Quick' tab. The 'Variables' section shows 'Dependent: none' and 'Independent: none'. The 'Options' section includes 'Weighted moments' (unchecked), 'DF = W-1 N-1', and 'MD deletion' (radio buttons for 'Casewise', 'Pairwise', and 'Mean substitution').

STATISTICA - Spreadsheet1

File Edit View Insert Format Statistics Graphs Tools Data Window Help

Add to Workbook Add to Report

Arial 10 B I U

Data: Spreadsheet1* (10v by 10c)

	1	2	3	4	5
	t	K	x	y	V
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	

Multiple Regression Results: Spreadsheet1

Multiple Regression Results

Dependent: y Multiple R = ,99369313 F = 628,2349
 R² = ,98742603 df = 1,8
 No. of cases: 10 adjusted R² = ,98885428 p = ,000000
 Standard error of estimate: ,286984631
 Intercept: 25,769691066 Std. Error: ,9517033 t(8) = 27,077 p = ,0000

x beta=-,99

(significant betas are highlighted)

Alpha for highlighting effects: ,05

Quick: Advanced Residuals/assumptions/prediction

Summary: Regression results Partial correlations
 ANOVA (Overall goodness of fit) Redundancy
 Covariance of coefficients Stepwise regression summary
 Current swage matrix ANOVA adjusted for mean

OK Cancel Options

STATISTICA - Workbook5* - [Regression Summary for Dependent Variable: y (Spreadsheet1)]

File Edit View Insert Format Statistics Graphs Tools Data Workbook Window Help

Add to Workbook Add to Report

Arial 10 B I U

Data: Spreadsheet* (10v by 10c)

	1	2	3	4	5	6	7	8	9	10
	t	K	x	y	Var5	Var6	Var7	Var8	Var9	Var10
1	20	0.3	0.003413	-1.20397						
2	30	0.8	0.0033	-0.22314						
3	40	1	0.003195	0						

Workbook5* - Regression Summary for Dependent Variable: y (Spreadsheet)

Regression Summary for Dependent Variable: y (Spreadsheet)

R = ,99369313 R² = ,98742603 Adjusted R² = ,98885428
 F(1,8) = 628,23 p = ,000000 Std. Error of estimate: ,286984631

	Beta	Std. Err.	B	Std. Err.	t	p-level
	of Beta			of Beta		
Intercept			25,771	0,9517	27,077	0,000000
x	-0,993693	0,039645	-7,974	0,81	-17,23	0,000000

Summary Statistics: DV: y (Spreadsheet1) Regression Summary for Dependent Variable: y (Spreadsheet1)

STATISTICA - Spreadsheet

File Edit View Insert Format Statistics Graphs Tools Data Window Help

10 Arial

Data: Spreadsheet1* (10v by 10c)

	1	2	3	4	5
	t	K	x	y	Var5
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	

Variable 5

Name: K0 Type: Double

Measurement Type: Auto Length: 8 MD code: -9999

Display format: General

Long name (label or formula with Functions): =EXP(25.77)

Labels: use any text. Formulas: use variable names or v1, v2, ..., v0 is case #. Examples: (a) = mean(v1:v3, sqrt(v7), AGE) (b) = v1+v2; comment (after.)

ok5* - Regression Summary for Dependent Variable: y (Spreadsheet1)

Multiple Regression (Spreadsheet1) Regression results dialog

Regression Summary Statistics

N=10

	Beta	Std Err. of Beta	B	Std Err. of B	t(8)	p-level
Intercept			25.77	0.9517	27.0774	0.000000
x	-0.993693	0.039645	-7974.89	318.1733	-25.0646	0.000000

STATISTICA - Spreadsheet

File Edit View Insert Format Statistics Graphs Tools Data Window Help

10 Arial

Data: Spreadsheet1* (10v by 10c)

	1	2	3	4	5	6
	t	K	x	y	K0	
1	20	0.3	0.003413	-1.20397	1.56E+11	
2	30	0.8	0.0033	-0.22314	1.56E+11	
3	40	1	0.003195	0	1.56E+11	
4	50	2.5	0.003096	0.916291	1.56E+11	
5	60	5	0.003003	1.609438	1.56E+11	
6	70	9	0.002915	2.197225	1.56E+11	
7	80	24	0.002833	3.178054	1.56E+11	
8	90	45	0.002755	3.806662	1.56E+11	
9	100	90	0.002681	4.49981	1.56E+11	
10	110	190	0.002611	5.247024	1.56E+11	

Variable 6

Name: E Type: Double

Measurement Type: Auto Length: 8 MD code: -9999

Display format: General

Long name (label or formula with Functions): =8.31*(-7974.89)

Labels: use any text. Formulas: use variable names or v1, v2, ..., v0 is case #. Examples: (a) = mean(v1:v3, sqrt(v7), AGE) (b) = v1+v2; comment (after.)

ok5* - Regression Summary for Dependent Variable: y (Spreadsheet1)

Multiple Regression (Spreadsheet1) Regression results dialog

Regression Summary Statistics

N=10

	Beta	Std Err. of Beta	B	Std Err. of B	t(8)	p-level
Intercept			25.77	0.9517	27.0774	0.000000
x	-0.993693	0.039645	-7974.89	318.1733	-25.0646	0.000000