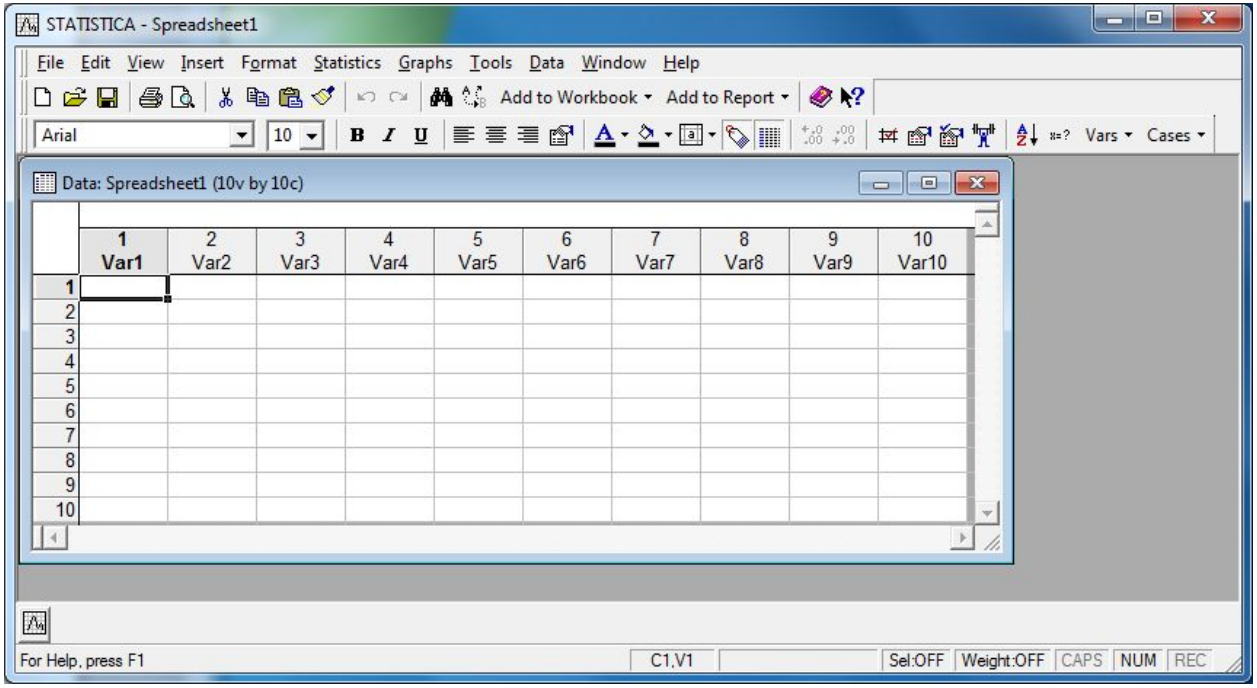
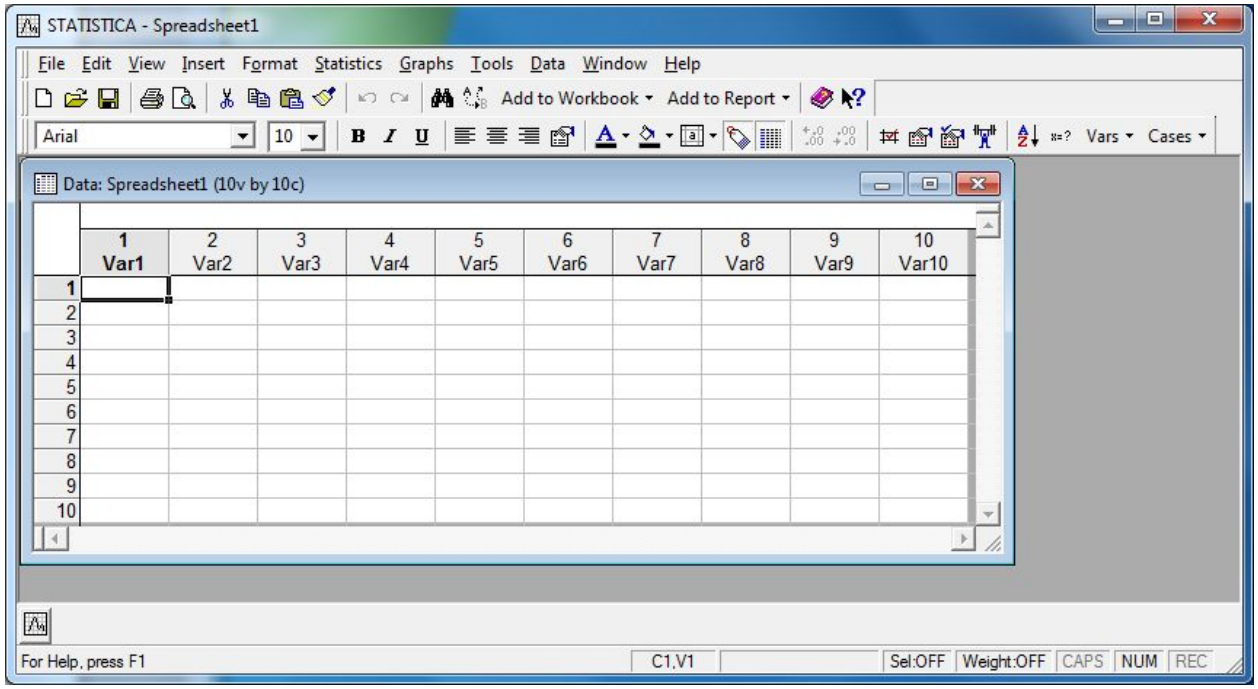
Статистический анализ данных в пакете STATISTICA

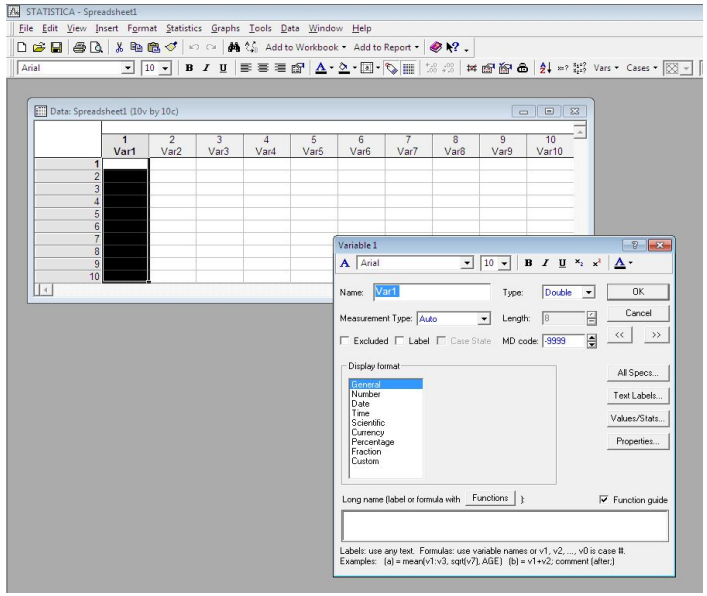
Задача. По экспериментальным значениям *константы скорости химической реакции* (K), полученным при различной *температуре* (t), *определить параметры уравнения Аррениуса* K_0 и E. Уравнение Аррениуса имеет следующий вид:

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

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







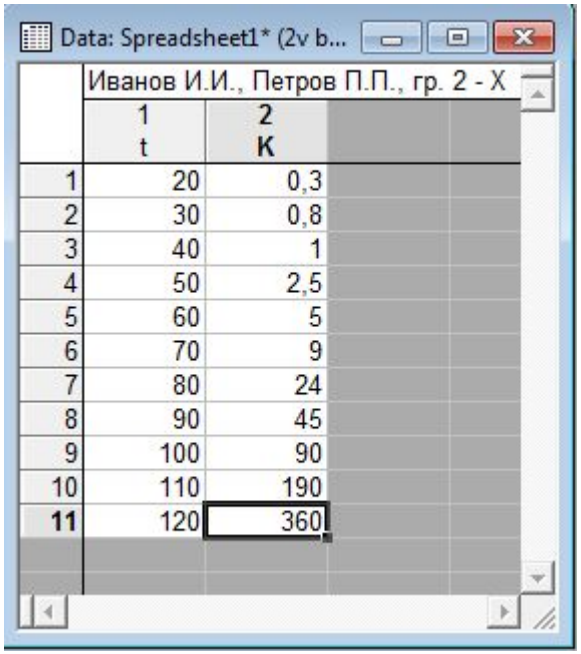
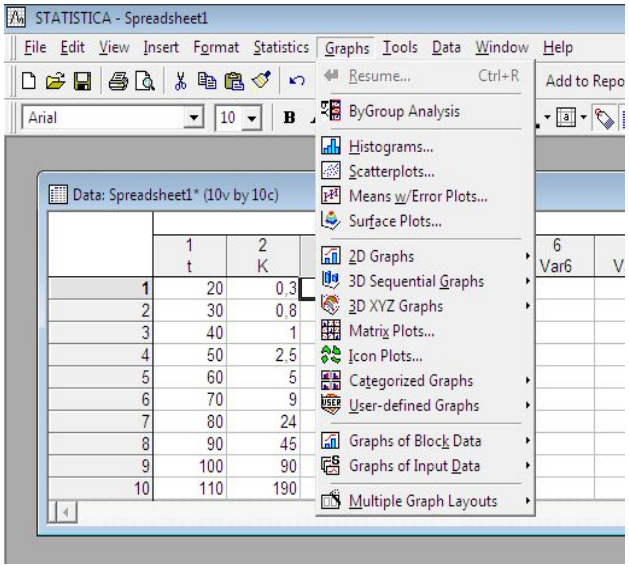


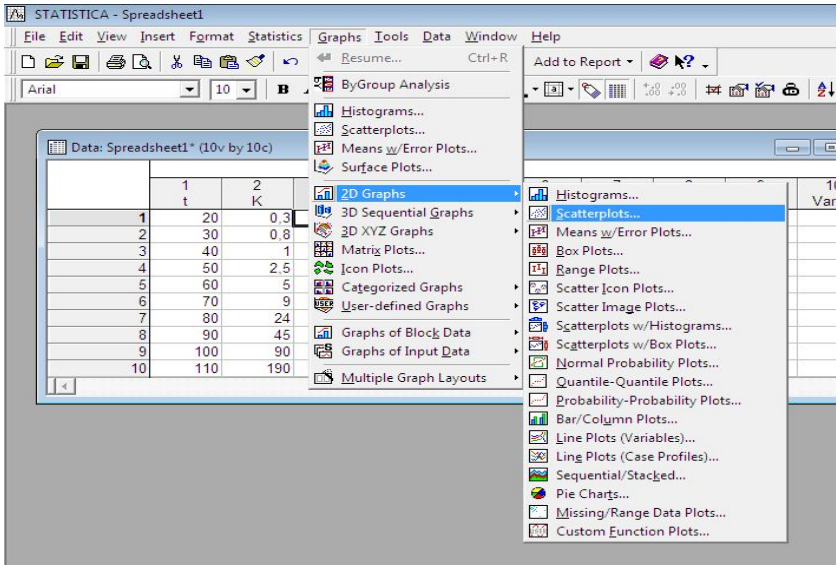
Рис. 1. Пример таблицы с экспериментальными данными

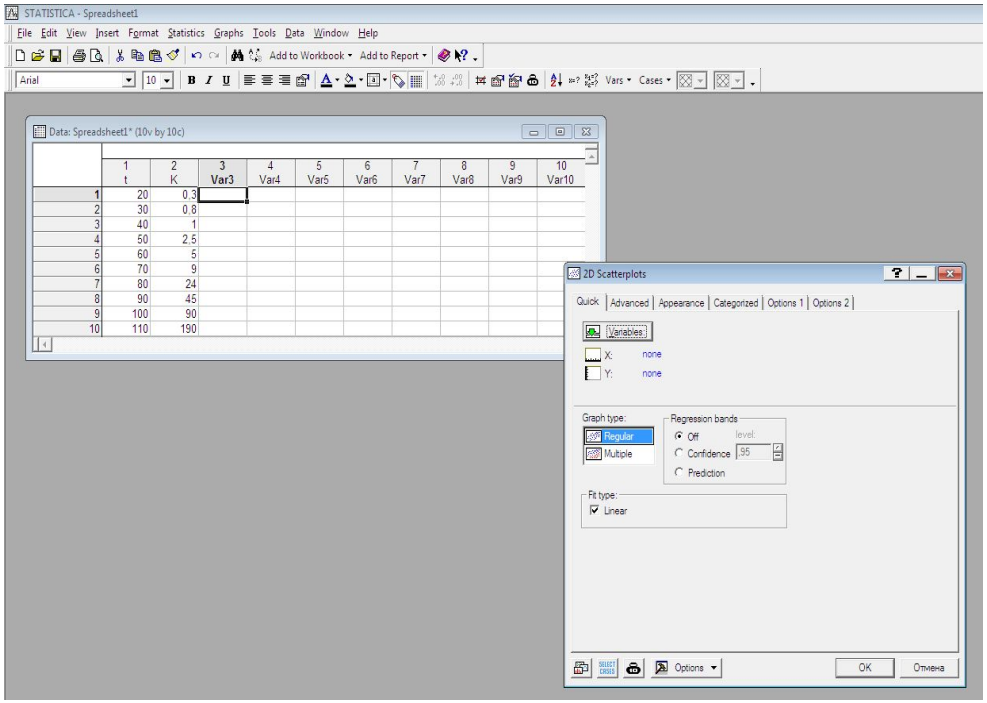
Визуализация исходных данных

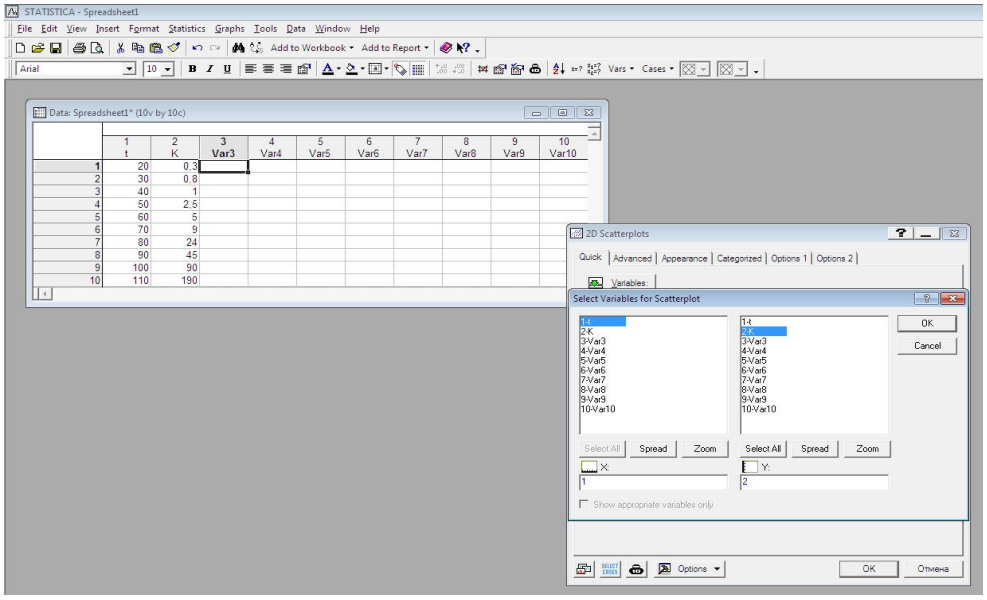
Вполне логично перед поиском зависимости сначала построить график зависимости по исходным данным. Это позволит предугадать общий вид искомой зависимости.

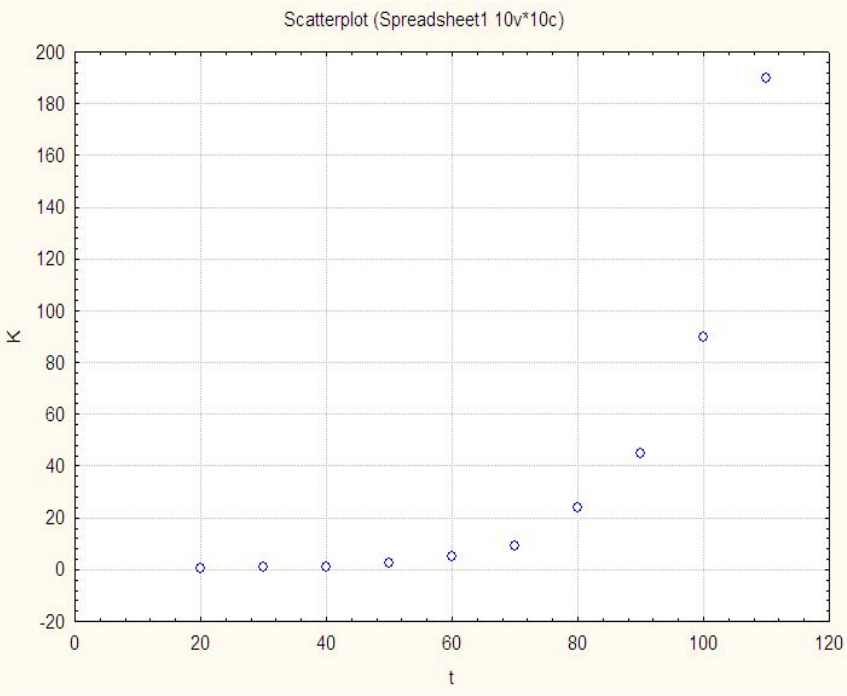
- Щелкаем ЛКМ в меню команд **Graph**, выбираем **Scatterplots**;
- В окне **2D Scatterplots** на закладку **Quick** щелкаем по кнопке **Variables**
- В Окне **Select Variables for Scatterplots** в левом списке (**X:**) выбираем переменную **t**, а правом (**Y:**) – **K**. Щелкаем ОК. Отключаем **Linear fit** и щелкаем по клавише окна **OK**.






STATISTICA - Spreadsheet1
File Edit View Insert Format Statistics Graphs Tools Data Window Help
Add to Workbook
Add to Report
Data: Spreadsheet1* (10v by 10c)
1 t 2 K 3 Var3 4 Var4 5 Var5 6 Var6 7 Var7 8 Var8 9 Var9 10 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
Multiple
Regression bands
Off
Confidence
Prediction
level: 95
Fit type:
Linear
Options
OK
Отмена


STATISTICA - Spreadsheet1
File Edit View Insert Format Statistics Graphs Tools Data Window Help
Add to Workbook
Add to Report
Data: Spreadsheet1* (10v by 10c)
1 t 2 K 3 Var3 4 Var4 5 Var5 6 Var6 7 Var7 8 Var8 9 Var9 10 Var10
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
OK
Cancel
Select All
Spread
Zoom
X:
Y:
Show appropriate variables only
Options
OK
Отмена

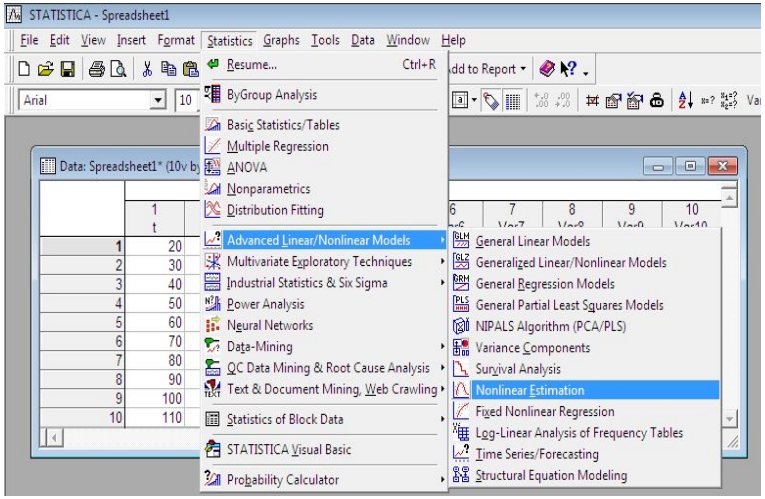

Scatterplot (Spreadsheet1 10v*10c)
K
t
200
180
160
140
120
100
80
60
40
20
0
-20
0
20
40
60
80
100
120

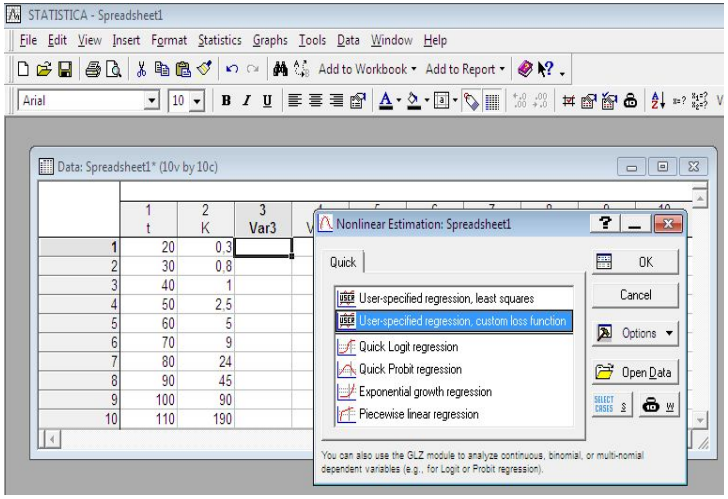
Модуль нелинейного оценивания

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

Вызов стартовой панели модуля

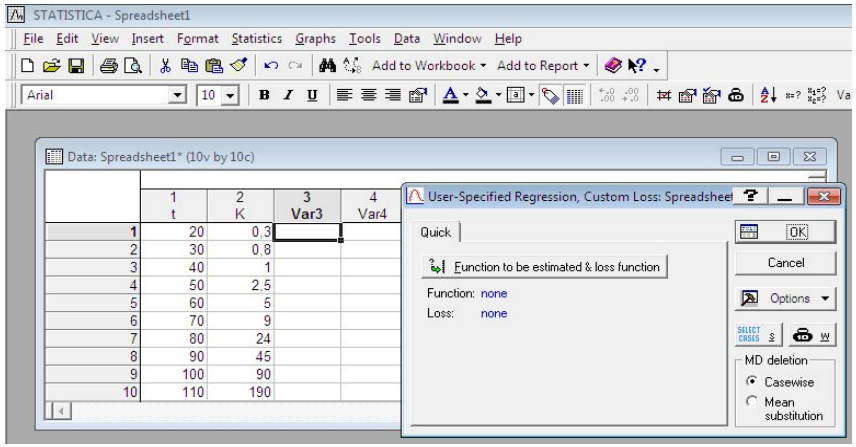
- Прежде всего, в пакете должна быть открыта таблица с исходными обрабатываемыми данными.
- Щелкаем меню команд **Statictics**, выбираем **Advanced Linear / Nonlinear Estimation**, в под меню щелкаем **Nonlinear Estimation**.
- В окно стартовой панели модуля выбираем **User-specified regression & custom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

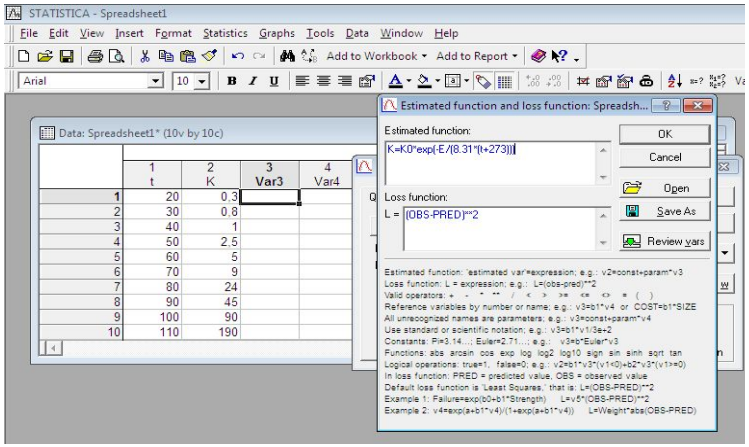


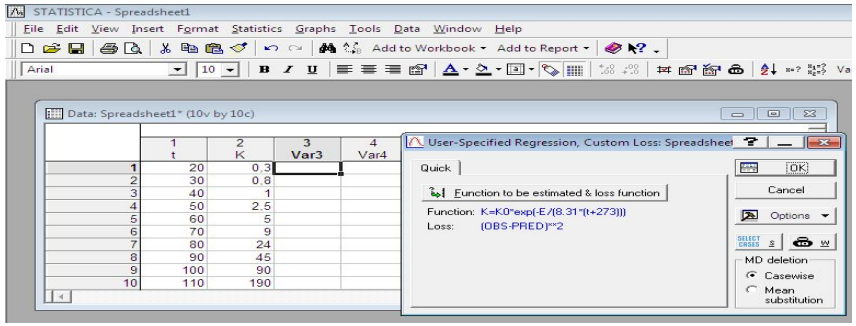


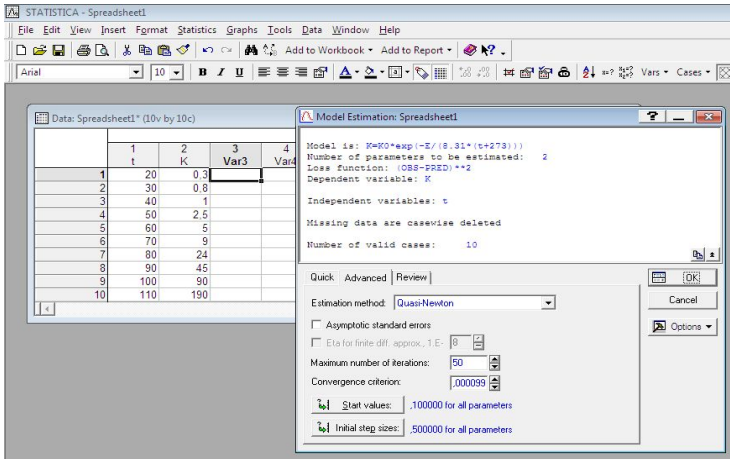
- В окне щелкаем кнопку $Functiontobeestimated\&lossfunction$, в поле **Estimated Function** вводим функцию в компьютерном виде:

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



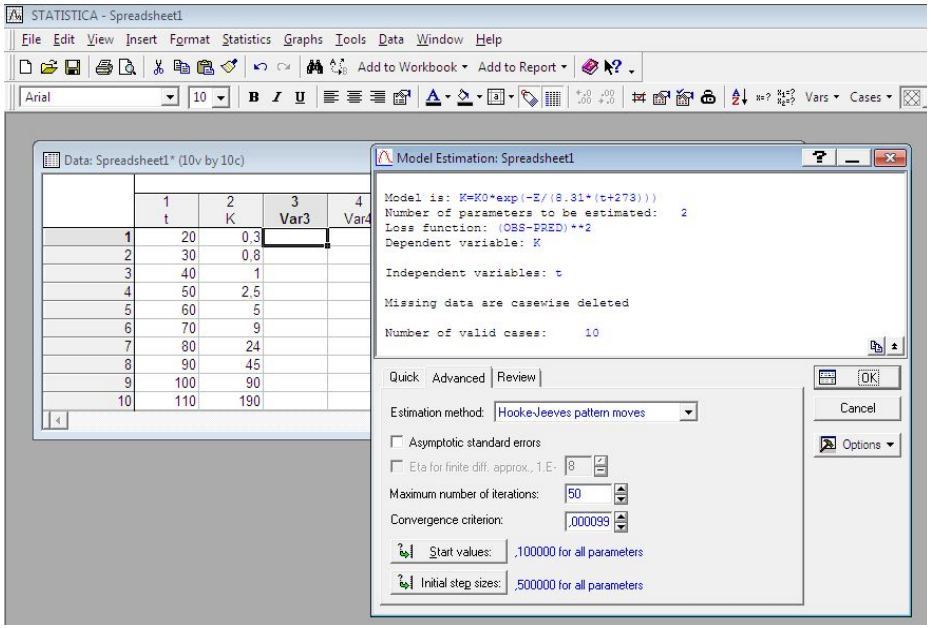


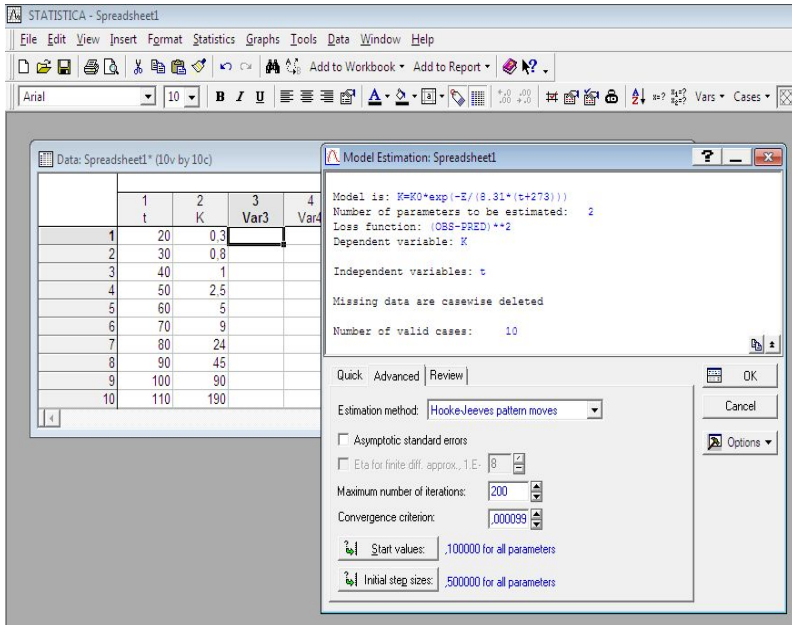


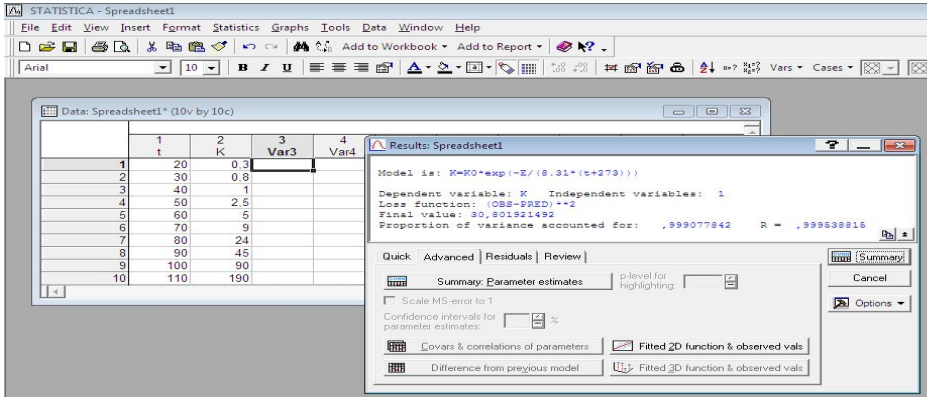


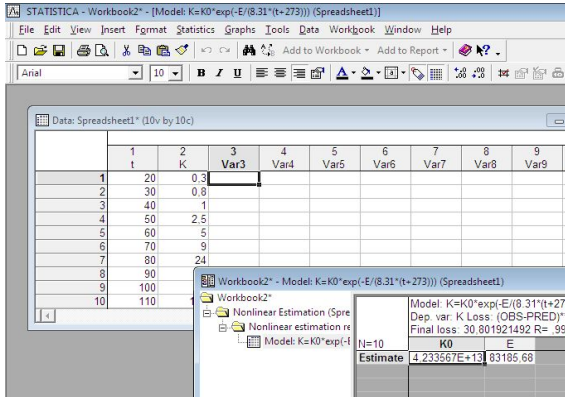
Выбор минимизируемой функции и численного метода поиска оценок:

- 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**). 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**.


STATISTICA - Spreadsheet1
File Edit View Insert Format Statistics Graphs Tools Data Window Help
Add to Workbook
Add to Report
Arial
10
Vars
Cases
Data: Spreadsheet1* (10v by 10c)
1 t
2 K
3 Var3
4 Var4
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
Model Estimation: Spreadsheet1
Model is: K=K0*exp(-E/(8,31*(t+273)))
Number of parameters to be estimated: 2
Loss function: (OBS-PRED)**2
Dependent variable: K
Independent variables: t
Missing data are casewise deleted
Number of valid cases: 10
Quick
Advanced
Review
OK
Cancel
Options
Estimation method:
Hooke-Jeeves pattern moves
Asymptotic standard errors
Eta for finite diff. approx., 1.E-
8
Maximum number of iterations:
50
Convergence criterion:
,000099
Start values:
,100000 for all parameters
Initial step sizes:
,500000 for all parameters


STATISTICA - Spreadsheet1
File Edit View Insert Format Statistics Graphs Tools Data Window Help
Add to Workbook
Add to Report
Arial
10
Vars
Cases
Data: Spreadsheet1* (10v by 10c)
1 t
2 K
3 Var3
4 Var4
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
Model Estimation: Spreadsheet1
Model is: K=K0*exp(-E/(8,31*(t+273)))
Number of parameters to be estimated: 2
Loss function: (OBS-PRED)**2
Dependent variable: K
Independent variables: t
Missing data are casewise deleted
Number of valid cases: 10
Quick
Advanced
Review
OK
Cancel
Options
Estimation method:
Hooke-Jeeves pattern moves
Asymptotic standard errors
Eta for finite diff. approx., 1.E-
8
Maximum number of iterations:
200
Convergence criterion:
,000099
Start values:
,100000 for all parameters
Initial step sizes:
,500000 for all parameters


STATISTICA - Spreadsheet1
File Edit View Insert Format Statistics Graphs Tools Data Window Help
Add to Workbook
Add to Report
Arial
10
Vars
Cases
Data: Spreadsheet1* (10v by 10c)
1 t
2 K
3 Var3
4 Var4
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
Results: Spreadsheet1
Model is: K=K0*exp(-E/(8,31*(t+273)))
Dependent variable: K Independent variables: 1
Loss function: (OBS-PRED)**2
Final value: 30,801921492
Proportion of variance accounted for: ,999077842 R = ,999538815
Quick
Advanced
Residuals
Review
Summary
Cancel
Options
Summary: Parameter estimates
p-level for highlighting:
Scale MS-error to 1
Confidence intervals for parameter estimates:
Covars & correlations of parameters
Fitted 2D function & observed vals
Difference from previous model
Fitted 3D function & observed vals


STATISTICA - Workbook2* - [Model: K=K0*exp(-E/(8.31*(t+273))) (Spreadsheet1)]
File Edit View Insert Format Statistics Graphs Tools Data Workbook Window Help
Add to Workbook
Add to Report
Arial
10
Data: Spreadsheet1* (10v by 10c)
1 t
2 K
3 Var3
4 Var4
5 Var5
6 Var6
7 Var7
8 Var8
9 Var9
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
9 100
10 110
Workbook2* - Model: K=K0*exp(-E/(8.31*(t+273))) (Spreadsheet1)
Workbook2*
Nonlinear Estimation (Spre
Nonlinear estimation re
Model: K=K0*exp(-
Model: K=K0*exp(-E/(8.31*(t+27
Dep. var: K Loss: (OBS-PRED)*
Final loss: 30,801921492 R= ,99
N=10
K0
E
Estimate
4,233567E+13
83185,68

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, … 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):

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

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:

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

где коэффициенты b_0 и b_1 рассчитываются из условия:

$\Phi = \sum_{i=1}^{m} (\widehat{Y_i} - Y_i)^2$ (4)

Разность $\widehat{Y_i} - Y_i$ называют остатком i-го опыта. По его величине можно судить о качестве подгонки линейно зависимости. Выборочная регрессия (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:

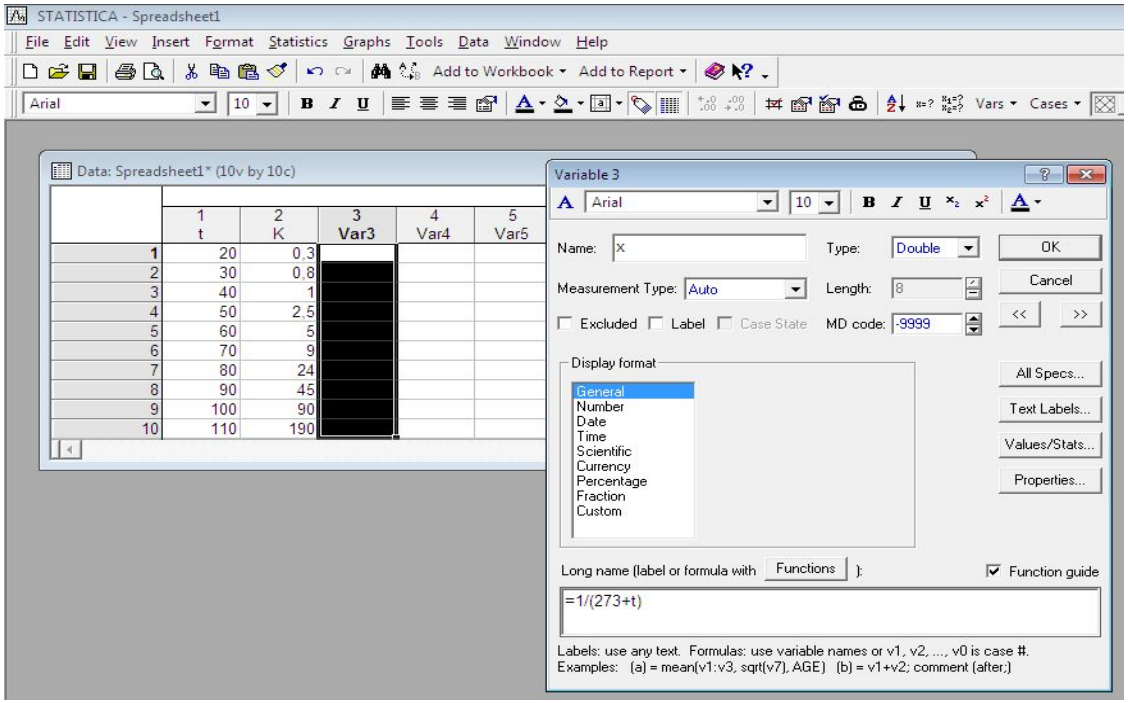
$\ln(K) = \ln(K_0) - \frac{E}{R} * \frac{1}{T}$ или в виде регрессии $Y = b_0 + b_1 * X$ (5)

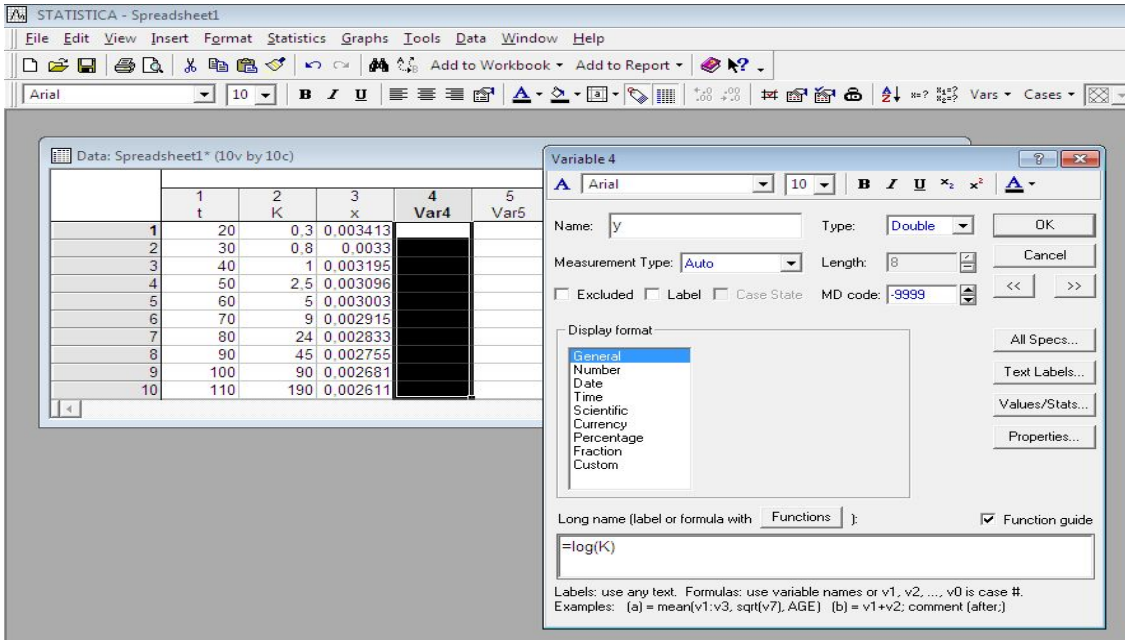
В таком случае для решения задачи необходимо найти значения коэффициентов линейной регрессии b_0 и b_1 и от них вернуться к исходным параметрам:

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

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

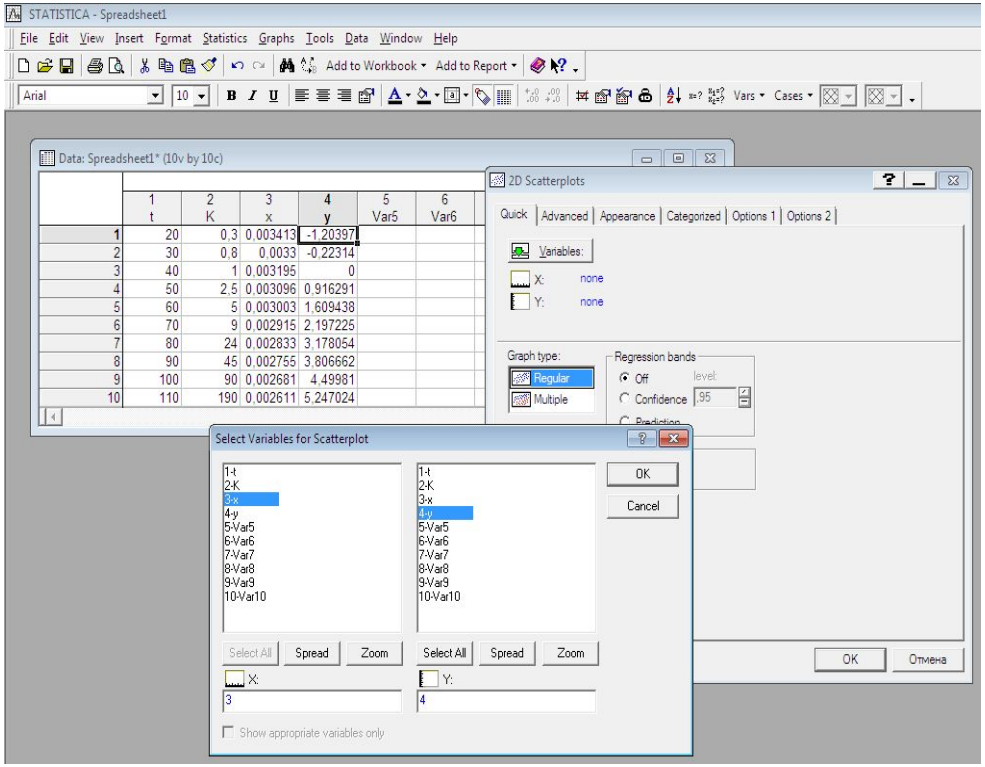
Обработка будет выполняться в модуле множественная регрессия (Multiple Regression).


STATISTICA - Spreadsheet1
File Edit View Insert Format Statistics Graphs Tools Data Window Help
Add to Workbook
Add to Report
Arial
10
Vars
Cases
Data: Spreadsheet1* (10v by 10c)
1 t
2 K
3 Var3
4 Var4
5 Var5
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
Variable 3
Name: x
Type: Double
OK
Cancel
Measurement Type: Auto
Length: 8
Excluded
Label
Case State
MD code: -9999
Display format
General
Number
Date
Time
Scientific
Currency
Percentage
Fraction
Custom
All Specs...
Text Labels...
Values/Stats...
Properties...
Long name (label or formula with Functions):
Function guide
=1/(273+t)
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;)

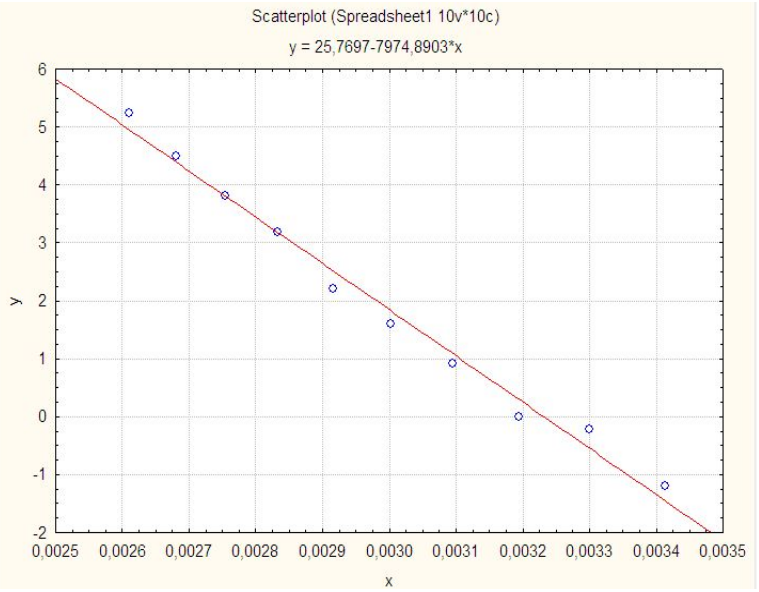

STATISTICA - Spreadsheet1
File Edit View Insert Format Statistics Graphs Tools Data Window Help
Add to Workbook
Add to Report
Arial
10
Vars
Cases
Data: Spreadsheet1* (10v by 10c)
1 t
2 K
3 x
4 Var4
5 Var5
1 20 0,3 0,003413
2 30 0,8 0,0033
3 40 1 0,003195
4 50 2,5 0,003096
5 60 5 0,003003
6 70 9 0,002915
7 80 24 0,002833
8 90 45 0,002755
9 100 90 0,002681
10 110 190 0,002611
Variable 4
Name: y
Type: Double
OK
Cancel
Measurement Type: Auto
Length: 8
Excluded
Label
Case State
MD code: -9999
Display format
General
Number
Date
Time
Scientific
Currency
Percentage
Fraction
Custom
All Specs...
Text Labels...
Values/Stats...
Properties...
Long name (label or formula with Functions):
Function guide
=log(K)
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;)

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**.

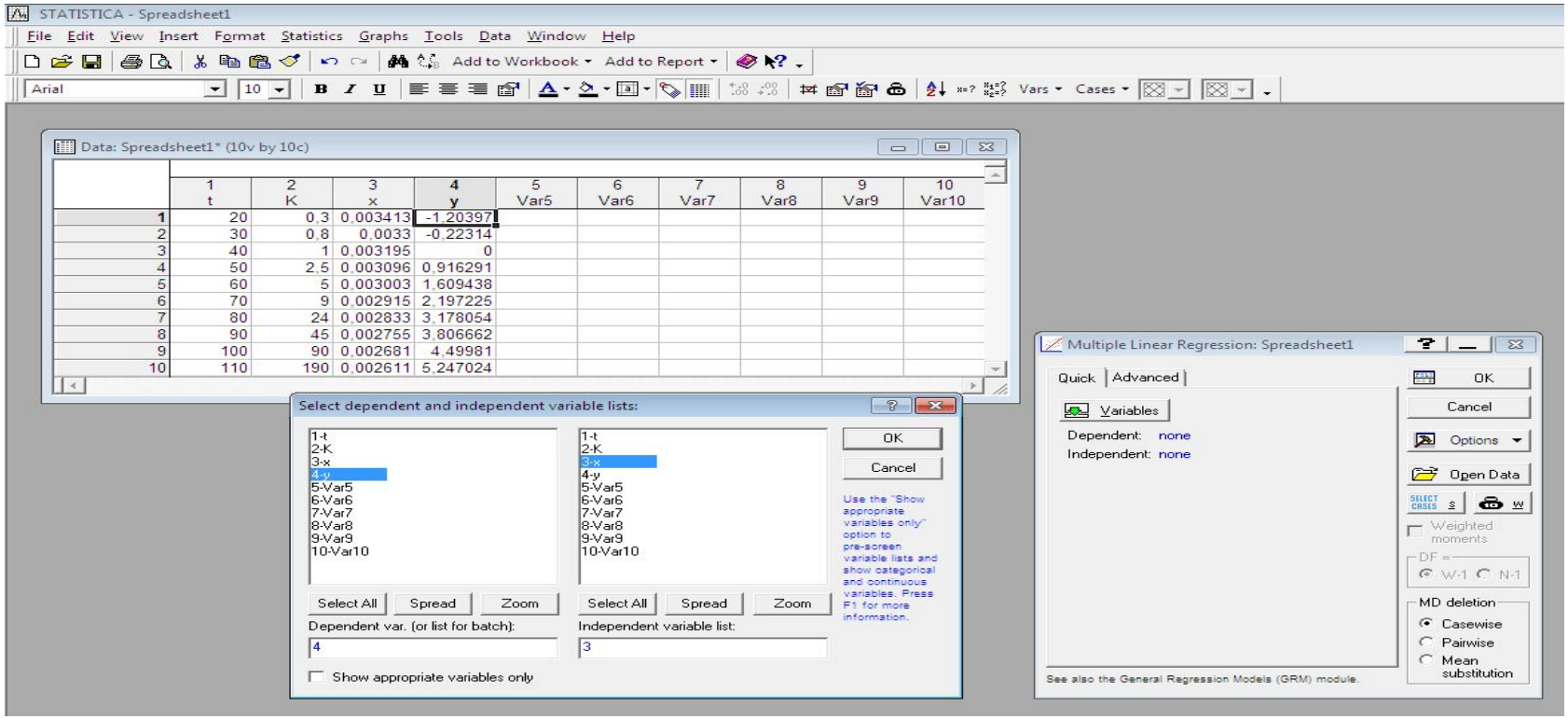


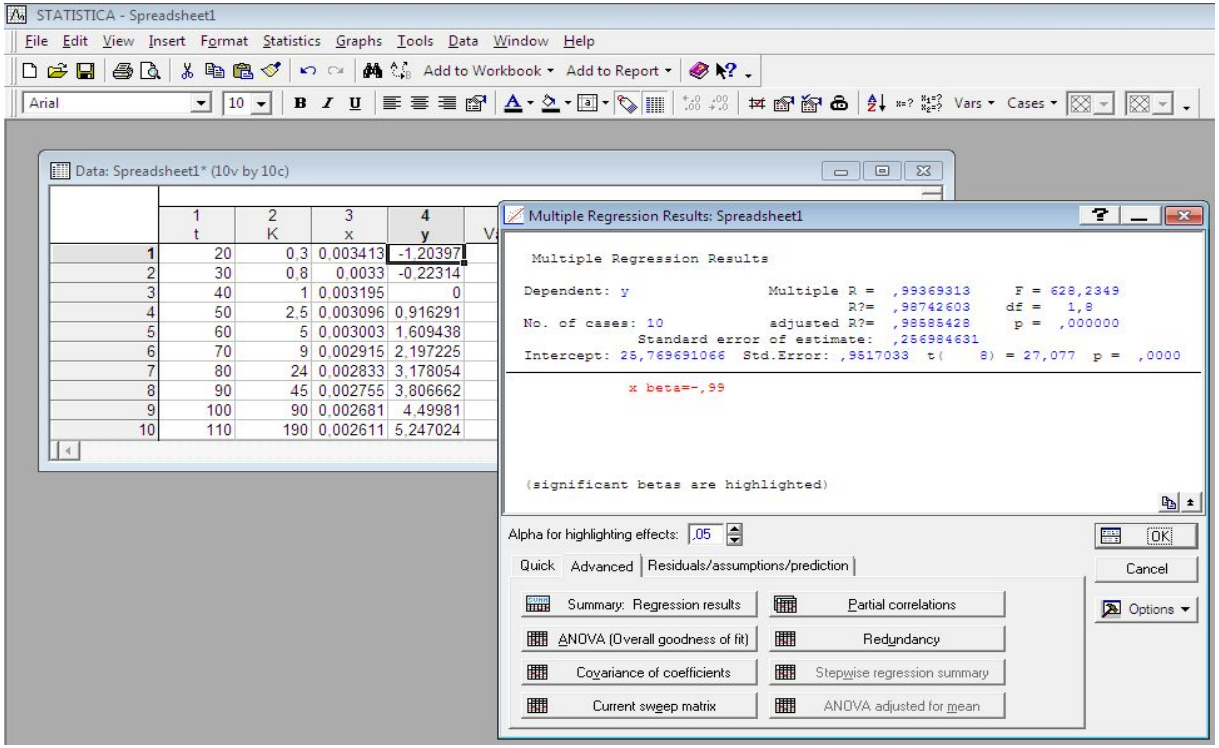
	1 t	2 K	3 x	4 y	5 Var5	6 Var6
1	20	0,3	0,003413	-1,20397		
2	30	0,8	0,0033	-0,22314		
3	40	1	0,003195	0		
4	50	2,5	0,003096	0,916291		
5	60	5	0,003003	1,609438		
6	70	9	0,002915	2,197225		
7	80	24	0,002833	3,178054		
8	90	45	0,002755	3,806662		
9	100	90	0,002681	4,49981		
10	110	190	0,002611	5,247024		

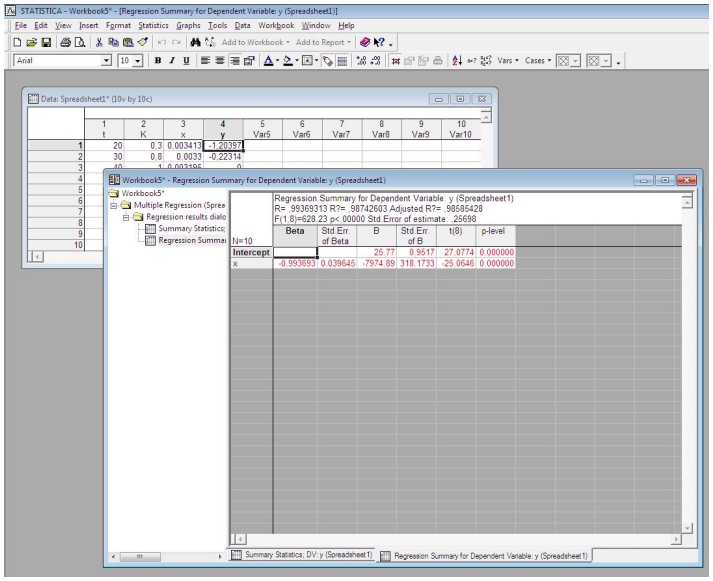


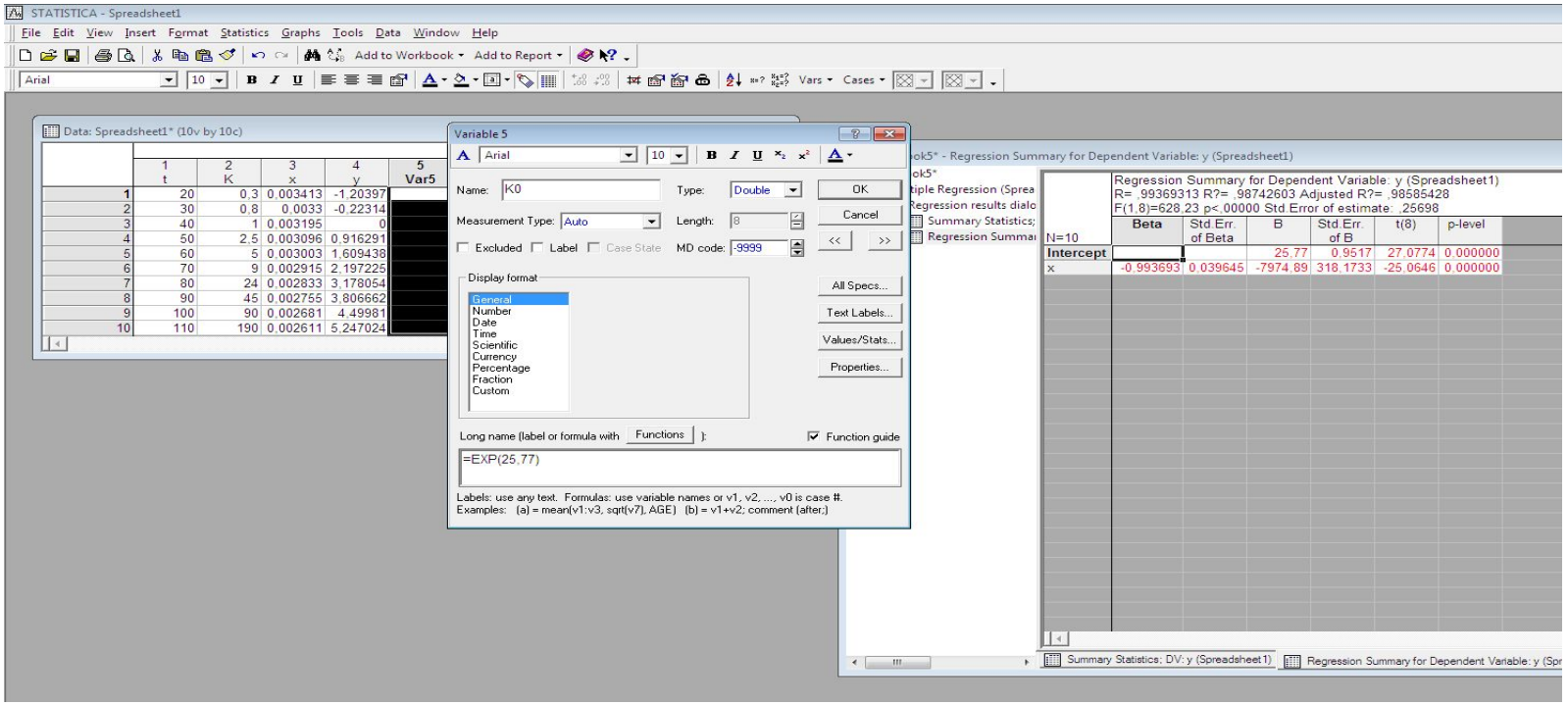
- В меню команд выбираем **Statistics**, далее – **Multiple Regression**
- В окне **Multiple Linear Regression** на закладке **Quick** щелкаем по кнопке **Variables**, в левом списке открывшегося окна ЛКМ выбираем зависимую (**Depended**) переменную – **Y**, а в правом – независимую (**Independed**)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

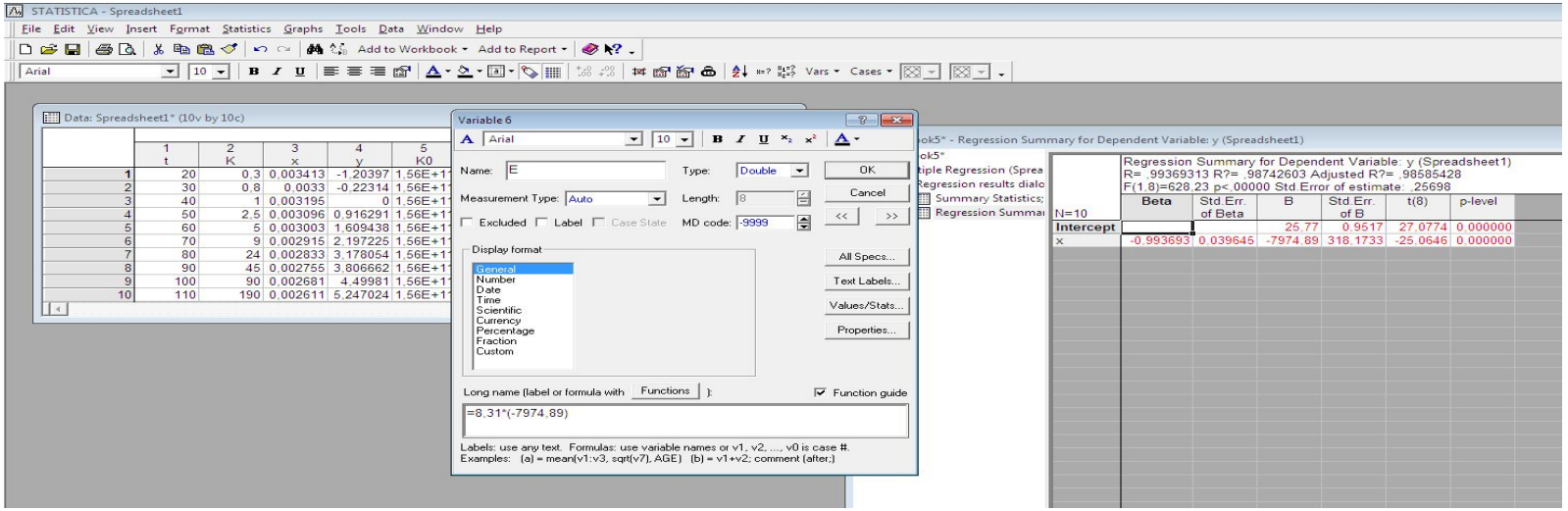
Для вывода результатов в **Multiple Linear Regression** щелкаем по кнопке **Summary: Regression Results**:




STATISTICA - Spreadsheet1
File Edit View Insert Format Statistics Graphs Tools Data Window Help
Add to Workbook
Add to Report
Arial
10
Vars
Cases
Data: Spreadsheet1* (10v by 10c)
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
Multiple Regression Results: Spreadsheet1
Multiple Regression Results
Dependent: y
Multiple R = ,99369313
R?= ,98742603
F = 628,2349
df = 1,8
No. of cases: 10
adjusted R?= ,98585428
p = ,000000
Standard error of estimate: ,256984631
Intercept: 25,769691066 Std.Error: ,9817033 t(8) = 27,077 p = ,0000
x beta=-,99
(significant betas are highlighted)
Alpha for highlighting effects: ,05
OK
Cancel
Options
Quick
Advanced
Residuals/assumptions/prediction
Summary: Regression results
Partial correlations
ANOVA (Overall goodness of fit)
Redundancy
Covariance of coefficients
Stepwise regression summary
Current sweep matrix
ANOVA adjusted for mean


STATISTICA - Workbook5* - [Regression Summary for Dependent Variable: y (Spreadsheet1)]
Workbook5* - Regression Summary for Dependent Variable: y (Spreadsheet1)
Regression Summary for Dependent Variable: y (Spreadsheet1)
R= .99369313 R?= .98742603 Adjusted R?= .98585428
F(1,8)=628.23 p< .00000 Std.Error of estimate: .25698
Beta
Std.Err. of Beta
B
Std.Err. of B
t(8)
p-level
N=10
Intercept
25.77
0.9517
27.0774
0.000000
x
-0.993693
0.039645
-7974.89
318.1733
-25.0646
0.000000
Summary Statistics: DV: y (Spreadsheet1)
Regression Summary for Dependent Variable: y (Spreadsheet1)


STATISTICA - Spreadsheet1
File Edit View Insert Format Statistics Graphs Tools Data Window Help
Add to Workbook
Add to Report
Data: Spreadsheet1* (10v by 10c)
Variable 5
Name: K0
Type: Double
OK
Cancel
Measurement Type: Auto
Length: 8
Excluded Label Case State
MD code: -9999
Display format
General
Number
Date
Time
Scientific
Currency
Percentage
Fraction
Custom
All Specs...
Text Labels...
Values/Stats...
Properties...
Long name (label or formula with Functions):
Function guide
=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;)
Regression Summary for Dependent Variable: y (Spreadsheet1)
R= ,99369313 R?= ,98742603 Adjusted R?= ,98585428
F(1,8)=628,23 p<,00000 Std.Error of estimate: ,25698
Beta
Std.Err. of Beta
B
Std.Err. of B
t(8)
p-level
N=10
Intercept
25,77
0,9517
27,0774
0,000000
x
-0,993693
0,039645
-7974,89
318,1733
-25,0646
0,000000
Summary Statistics; DV: y (Spreadsheet1)
Regression Summary for Dependent Variable: y (Spreadsheet1)


STATISTICA - Spreadsheet1
File Edit View Insert Format Statistics Graphs Tools Data Window Help
Add to Workbook
Add to Report
Data: Spreadsheet1* (10v by 10c)
Variable 6
Name: E
Type: Double
OK
Cancel
Measurement Type: Auto
Length: 8
Excluded Label Case State
MD code: -9999
Display format
General
Number
Date
Time
Scientific
Currency
Percentage
Fraction
Custom
All Specs...
Text Labels...
Values/Stats...
Properties...
Long name (label or formula with Functions):
Function guide
=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;)
Regression Summary for Dependent Variable: y (Spreadsheet1)
R= ,99369313 R?= ,98742603 Adjusted R?= ,98585428
F(1,8)=628,23 p<,00000 Std.Error of estimate: ,25698
Beta
Std.Err. of Beta
B
Std.Err. of B
t(8)
p-level
N=10
Intercept
25,77
0,9517
27,0774
0,000000
x
-0,993693
0,039645
-7974,89
318,1733
-25,0646
0,000000