

## ЛАБОРАТОРНАЯ РАБОТА № 9 РЕГРЕССИЯ (ПРОДОЛЖЕНИЕ)

### НЕЛИНЕЙНАЯ РЕГРЕССИЯ

**Задание.** В таблице приведены данные года и численность населения (млн чел.). В 1965 году И.С. Шкловский получил гиперболический закон изменения численности населения Земли:  $Y = A/(B - Year)$ . Требуется получить закон Шкловского.

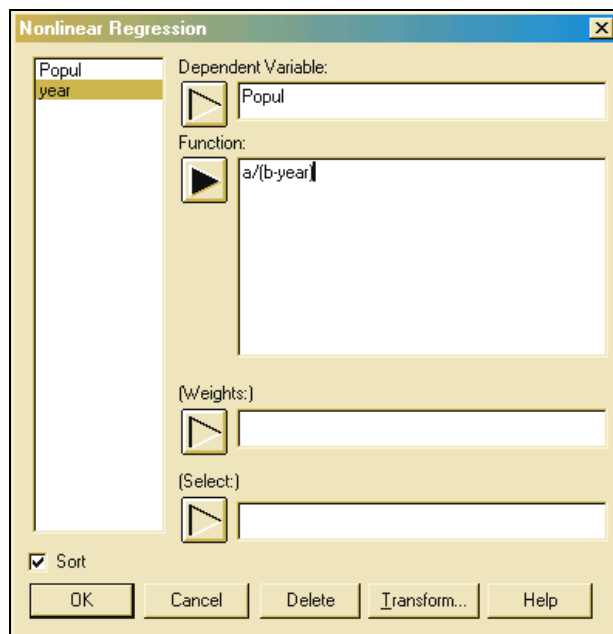
Введите данные в соответствии с таблицей:

Год (Year)	Численность (Popul) (млн чел)
1600	486
1650	545
1700	617
1750	728
1800	906
1850	1171
1900	1608
1920	1861
1930	2070
1940	2295
1950	2517
1960	3010

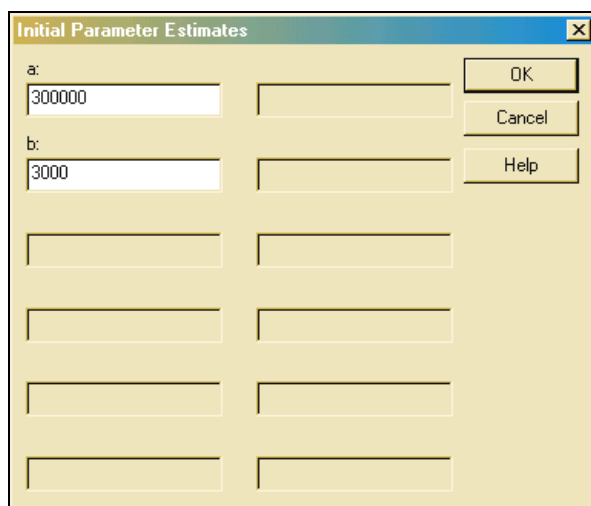
Данные о численности населения Земли по годам представляют собой **временной ряд**. При выявлении **тренда** временного ряда (его неслучайной составляющей) так же, как и в регрессионном анализе при выявлении связи двух случайных признаков, обычно используют один и тот же метод — метод наименьших квадратов.

В *Statgraphics 'e* есть возможность попробовать построить любую нелинейную регрессионную модель. В строке меню выберите **Relate**, в раскрывшемся

меню выберите **Multiple Factors**, а затем **Nonlinear Regression**. Откроется диалоговое окно **Nonlinear Regression**.

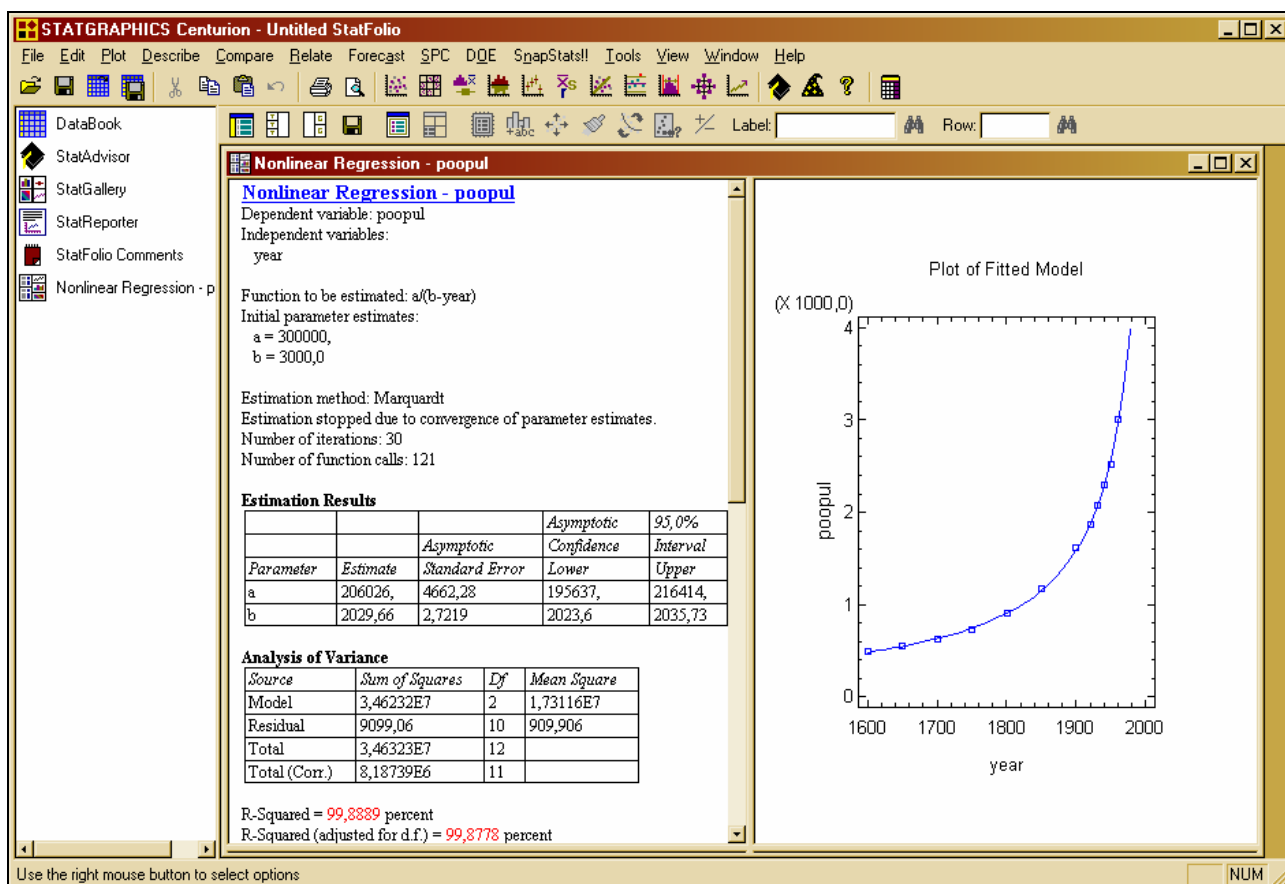


В этом окне надо выделить **Popul**, затем щелкнуть по стрелке в поле **Dependent Variable** (зависимая переменная), затем в поле **Function** надо задать формулу, коэффициенты которой вы хотите получить. В нашем случае это  $A/(B-Year)$ . Нажмите кнопку ОК. Перед вами раскроется окно.



В этом окне надо задать начальные параметры нелинейного регрессионного анализа. Как вам должно быть известно, не из всякой начальной точки метод будет сходиться к реальному результату (точке глобального минимума) и для ускорения сходимости лучше начинать двигаться (если это возможно) из более близкой к результату точке. Введите в поле  $a$  — 300000, а в поле  $b$  — 3000. Нажмите ОК.

## Раскрылось окно сводки первичного анализа

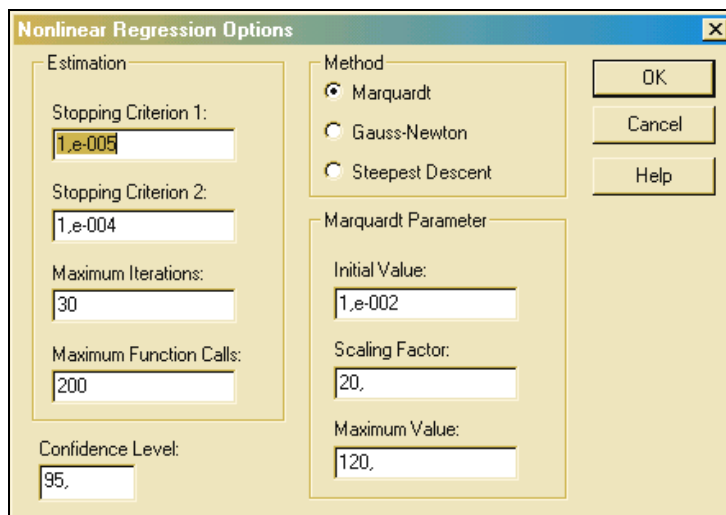


Прочитайте внимательно результаты, представленные в этом окне. Можете также пролистнуть и посмотреть результаты, написанные в StatAdvisor. Можно видеть, что в результате нелинейной модели регрессии мы получили уравнение  $Y = 206026,0 / (2029,66 - X)$

$R$ -квадрат ( $R$ -squared) указывает, что приспособленная модель объясняет 99,8889 % изменчивости в *Popul*. Скорректированный  $R$ -квадрат является более пригодным для сравнения моделей с другими количествами независимых переменных — 99,8778 %. **Стандартная ошибка оценки** (*Standard Error of Est*) показывает среднеквадратичное отклонение, равное 30,1646. Эта величина может использоваться для задания пределов прогнозирования новых наблюдений. **Средняя абсолютная ошибка** (*Mean absolute Error*) является средней величиной остатков и равна 18,5651. В *StatAdvisor* указано также, что уровень достоверности для неизвестных переменных — 95,0 %.

В процессе решения было выполнено 30 итераций, совершено 121 обращение к функции. В результате получилась необходимая сумма квадратов, дос-

тигнувшая минимума. Если в процессе вычислений необходимая точность не была достигнута, можно увеличить количество итераций. Щелкните правой кнопкой в окне, выберите *Analysis Options*, раскроется окно *Nonlinear Regression Options*, в котором можно выбрать нужные вам параметры.

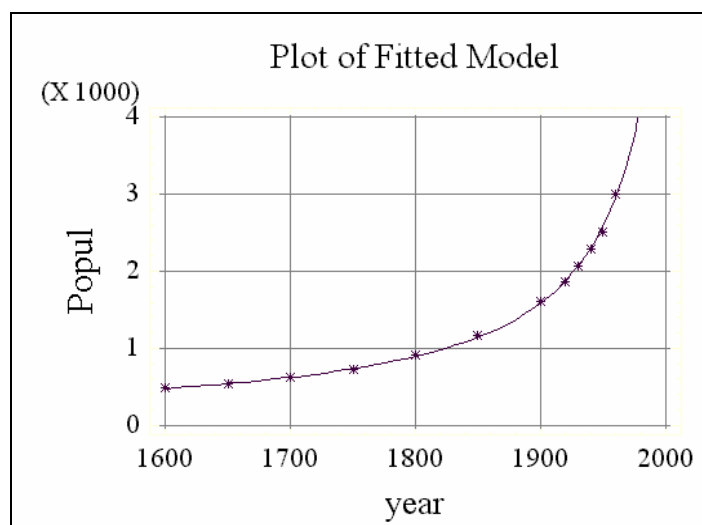


The dialog box titled "Nonlinear Regression Options" contains the following settings:

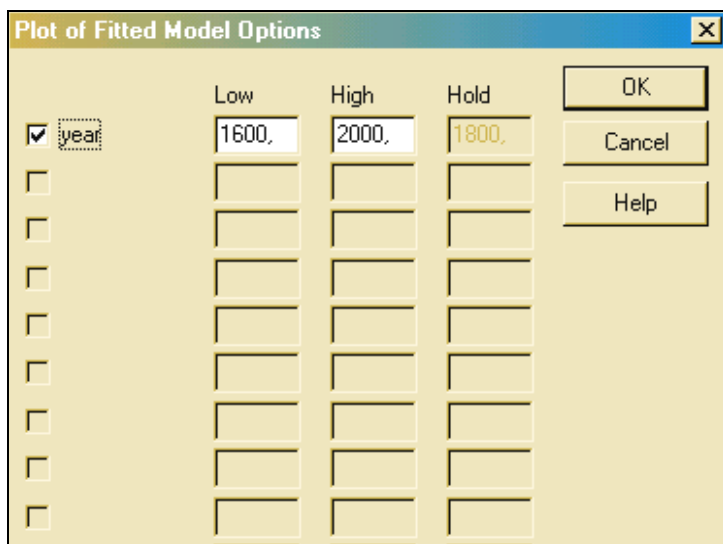
- Estimation:**
  - Stopping Criterion 1:  $1.e-005$
  - Stopping Criterion 2:  $1.e-004$
  - Maximum Iterations: 30
  - Maximum Function Calls: 200
  - Confidence Level: 95
- Method:**
  - ☒ Marquardt
  - ☐ Gauss-Newton
  - ☐ Steepest Descent
- Marquardt Parameter:**
  - Initial Value:  $1.e-002$
  - Scaling Factor: 20
  - Maximum Value: 120

Buttons: OK, Cancel, Help.

График показывает, что точки достаточно хорошо ложатся на кривую.



Нехорошо то, что по оси  $X$  график заканчивается 2000 годом, а из формулы видно, что 2030 год является особой точкой. Изменим настройки графика. Для этого щелкните правой кнопкой по графику, выберите *Pane Options*. Раскроется окно настроек.



В этом окне значение поля **High** измените с 2000 на 2050, нажмите кнопку ОК. На изменившемся графике видно, что около значения 2030 проходит асимптота.

1. Подумайте над получившимся результатом. Какое население Земли предсказывает эта формула к 2030 г?
2. Как можно было свести модель Шкловского к линейной?
3. Получите параметры модели Шкловского, используя результаты предыдущей лабораторной работы.

## ПОЛИНОМИАЛЬНАЯ РЕГРЕССИЯ

Процедура полиномиальной регрессии позволяет находить аналитические выражения связи двух переменных  $Y$  и  $X$  в виде степенного полинома

$$Y = a_0 + a_1X^1 + a_2X^2 + \dots + a_nX^n.$$

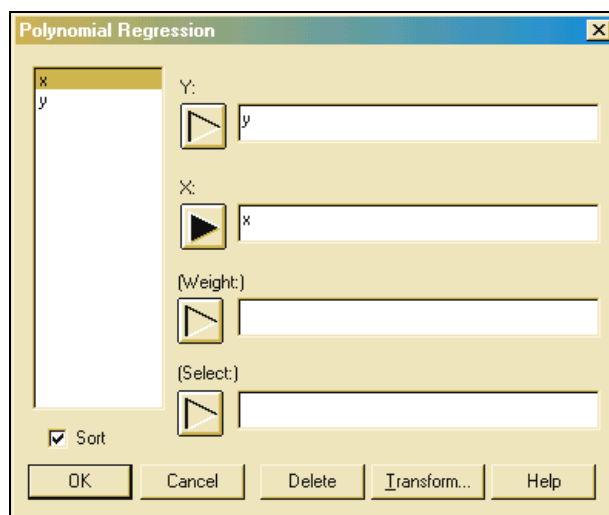
**Statgraphics** предоставляет возможность строить такие полиномы вплоть до **восьмой** степени.

**Задача.** Для установления зависимости между ежегодным потреблением бананов и уровнем годового дохода опрошена группа людей, полученные данные приведены в таблице.

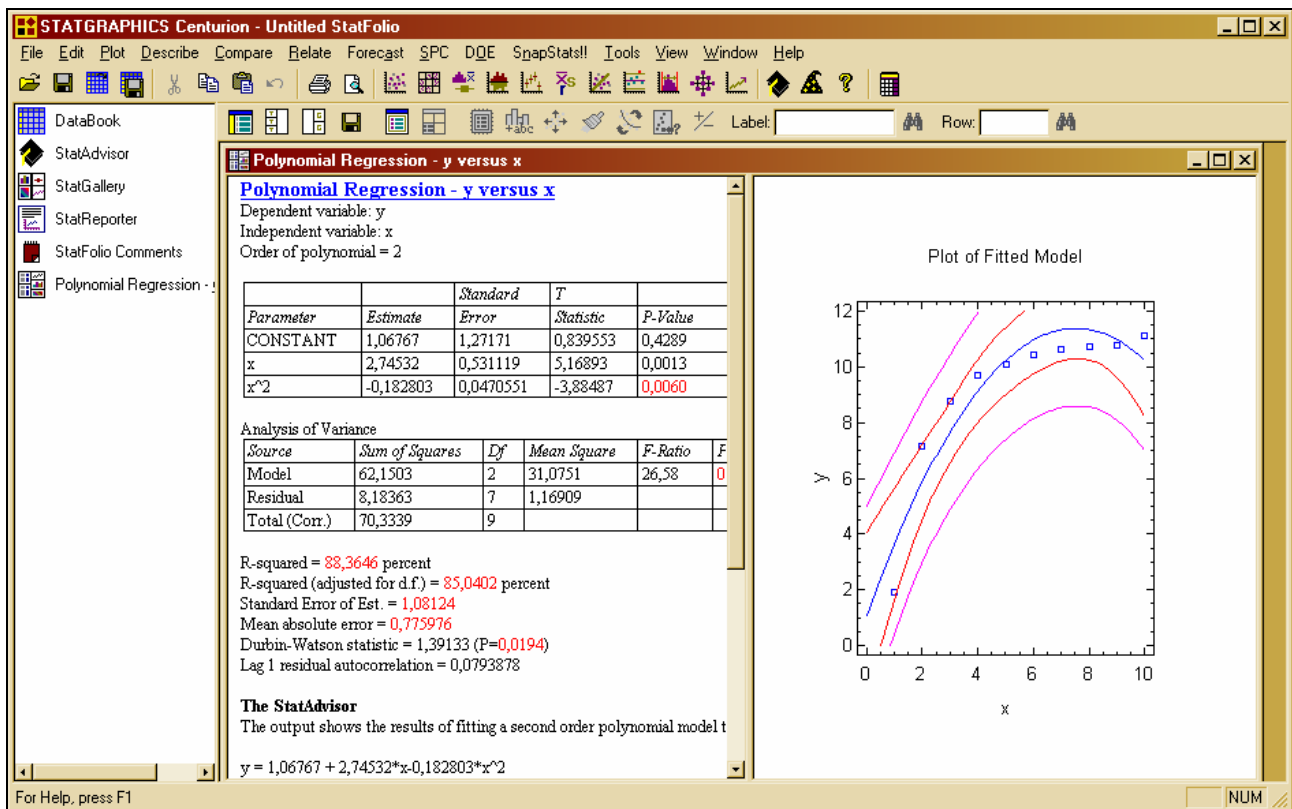
$Y$ — потребление бананов (в фунтах)	$X$ — доход (в 1000 дол)
1,93	1
7,13	2
8,78	3
9,69	4
10,09	5
10,42	6
10,62	7
10,71	8
10,79	9
11,13	10

Введите данные в соответствии с таблицей.

В строке меню выберите **Relate**, в раскрывшемся меню выберите **One Factor, Polynomial Regression**. Нажмите ОК. Раскроется окно, в котором в поле  $X$ : надо ввести  $x$ , в поле  $Y$ : —  $y$ .

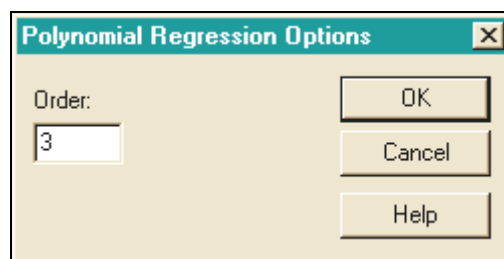


Нажмите кнопку ОК. Раскроется окно.

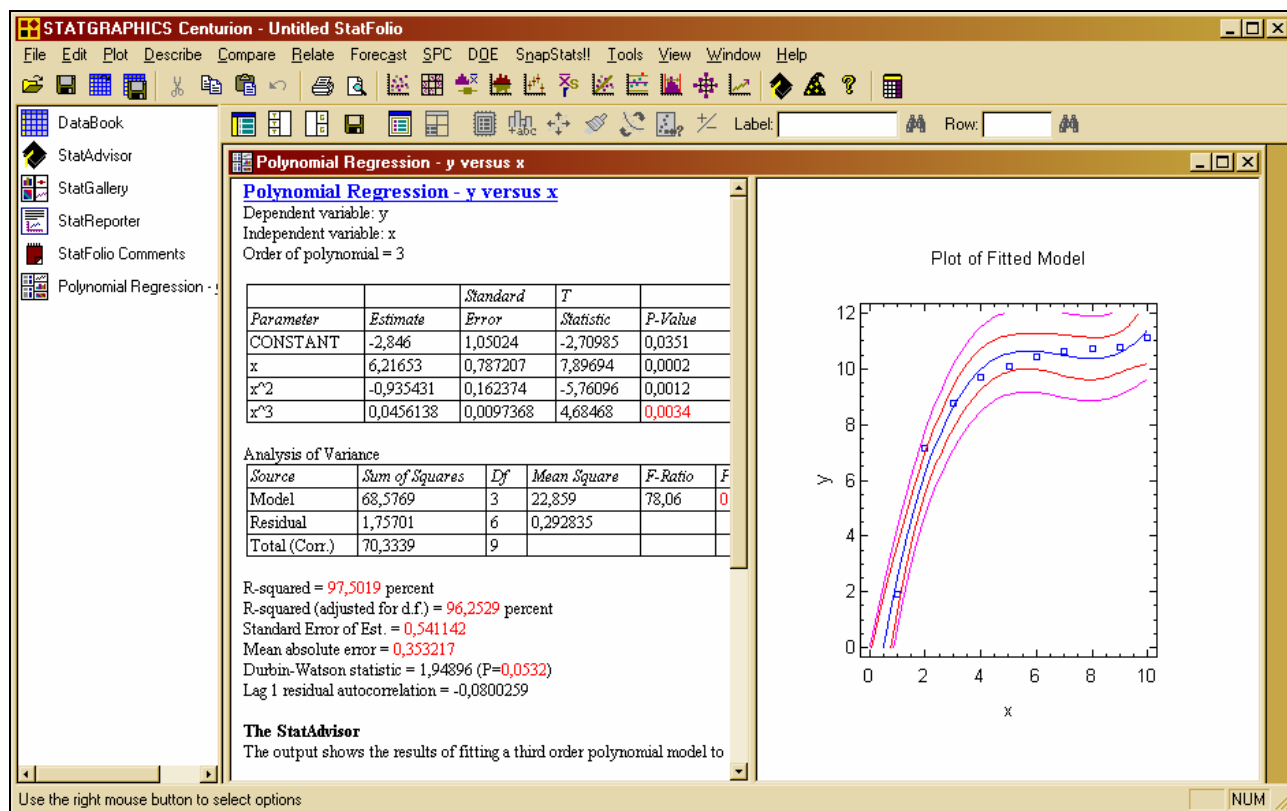


Перед вами сводка построенной модели регрессии второго порядка. Как следует из сводки, получена достаточно неплохая регрессионная модель. Об этом свидетельствует достаточно высокий коэффициент детерминации  $R$ -квадрат (85,04 %), низкое  $p$ -value (0,0005) по результатам дисперсионного анализа модели (*Analysis of Variance*) и другие показатели. Правда, к полученным результатам следует относиться осторожно из-за малого объема выборки.

Посмотрим теперь на графическое представление результатов. Графически модель второго порядка выглядит вполне удовлетворительно, точки попадают внутрь 95 % доверительной области. Построим полиномиальную регрессионную модель более высокого порядка. Для этого щелкните правой кнопкой по сводке результатов, в контекстном меню выберите **Analysis Options**:



В раскрывшемся меню задайте порядок 3, нажмите ОК. Появится новая сводка результатов



Видно, что модель третьего порядка обладает лучшими статистическими свойствами, чем модель первого порядка. Проанализируйте самостоятельно полученные результаты. Самостоятельно постройте график, попробуйте построить модели более высокого порядка. Обращайте внимание на *p-value* коэффициентов регрессии, следите за тем, чтобы они были статистически значимы. Результаты покажите преподавателю.

## ПОШАГОВАЯ МНОЖЕСТВЕННАЯ РЕГРЕССИЯ

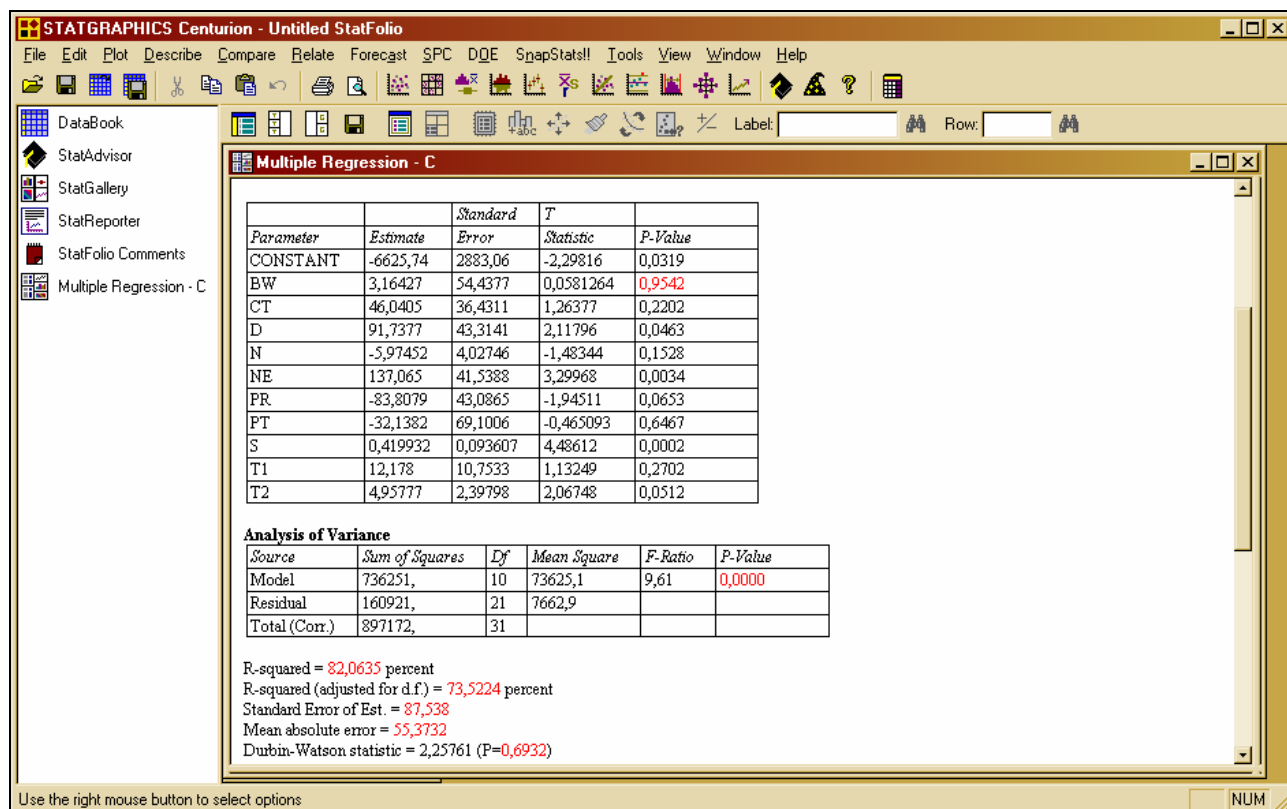
**Задание.** Скопируйте себе файл *Atomst.sf*. В таблице, находящейся в этом файле, приведены данные о капитальных затратах на строительство атомных электростанций с реактором водяного охлаждения. Данные собраны для 32 различных станций США. Требуется предсказать величину капитальных затрат на строительство новой станции, попробовать выделить наиболее значимые величины, влияющие на цену станции. Здесь  $C$  — цена в млн дол., приведенная к курсу 1976 года,  $D$  — срок разрешения на строительство;  $T1$  — время между обращением и получением разрешения на строительство;  $T2$  — время между получением оперативной лицензии и разрешением на строительство;  $S$  — но-



минальная мощность электростанции, Мвт; *PR* — наличие в той же самой местности ранее построенной электростанции на *PBO* (если значение равно 1, то имеется уже построенная станция); *NE* — характеристика района, в котором строится станция; *CT* — использование нагревательной башни (если значение равно 1, то используется, если 0 — нет); *BW* — использование силовой установки производства фирмы *Wilcox* (если значение равно 1, то используется, 0 — нет); *N* — суммарное количество электростанций, построенное архитектором-инженером станции; *PT* — электростанции, строящиеся под частичным надзором (1 если надзор есть, 0 — если нет).

В этой задаче зависимая переменная — цена станции, а независимые — *D, T1, T2, S, PR, NE, CT, BW, N, PT* — все остальные переменные, перечисленные в таблице. Зависимость между переменными предполагается линейной.

Прежде всего, введите данные в соответствии с таблицей. Затем в строке меню выберите **Relate**, в раскрывшемся контекстном меню выберите **Multiple Factors, Multiple Regression**. Раскроется окно, в нем в поле **Dependent Value** поместите переменную *C*, а все остальные переменные — в поле **Independent Value**. Вспоминайте предыдущую лабораторную работу, там вы выполняли примерно такое же задание. Нажмите ОК, перед Вами раскроется окно сводки проведенного анализа.



Почитайте *StatAdvisor*. В нем говорится, что построена модель

$$C = -6625,74 + 91,7377 \cdot D + 12,178 \cdot T1 + 4,95777 \cdot T2 + 0,419932 \cdot S -$$

$$3,8079 \cdot PR + 137,065 \cdot NE + 46,0405 \cdot CT + 3,16427 \cdot BW - 5,97452 \cdot N - 32,1382 \cdot PT.$$

Отмечается, что взаимоотношения переменных являются статистически значимыми на 99 % доверительном уровне. *R*-квадрат указывает, что модель отражает 82,0635 % изменчивости переменной *C*, а скорректированный *R*-квадрат с учетом степеней свободы (что является более подходящим для сравнения моделей с разными количествами переменных) составляет 73,5224 %. Стандартная ошибка равна 87,538 % и ее можно использовать в задании границ предсказания для новых наблюдений. Средняя абсолютная ошибка, представляющая собой среднюю величину остатков, составляет 55,3732. *StatAdvisor* предлагает уменьшить число переменных и исключить переменную **BW**, т. к. у этой переменной очень большое значение *p-value*.

Нажмите кнопку **Input Dialog** (самая левая кнопка в окне), появится окно для ввода данных, в поле **Independent Value** выберите переменную **BW**, которую хотите исключить, нажмите кнопку **Delete**. Нажмите ОК.

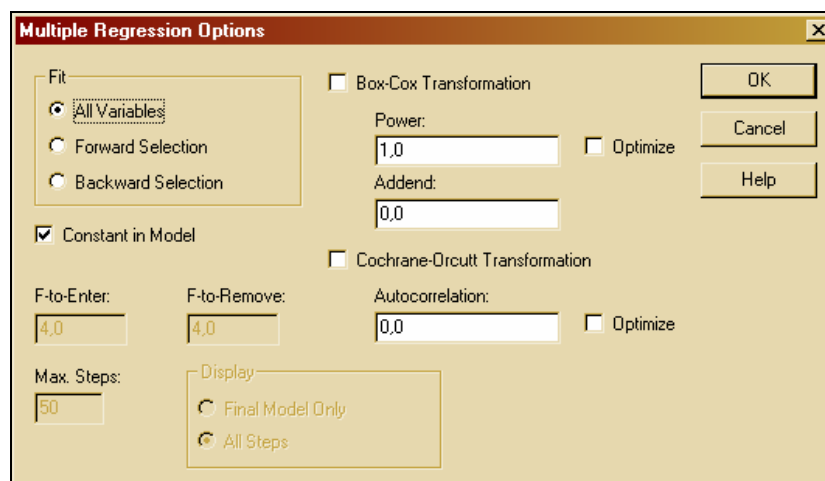
Таблица анализа сразу же изменится. Прочитайте самостоятельно результаты. *StatAdvisor* предлагает исключить переменную **PT**. Исключите ее самостоятельно. Далее исключите последовательно переменные **T1**, **CT**. *StatAdvisor* сообщает, что самая значимая независимая переменная — **PR**, с уровнем достоверности 95 %.

Исключите из рассмотрения переменную **PR**, уровень достоверности упадет до 90 %, *StatAdvisor* предлагает теперь исключить из рассмотрения и переменную **T2**. После этого уровень достоверности станет равным 95 % и значимой станет переменная **N**. Исключите и ее из рассмотрения.

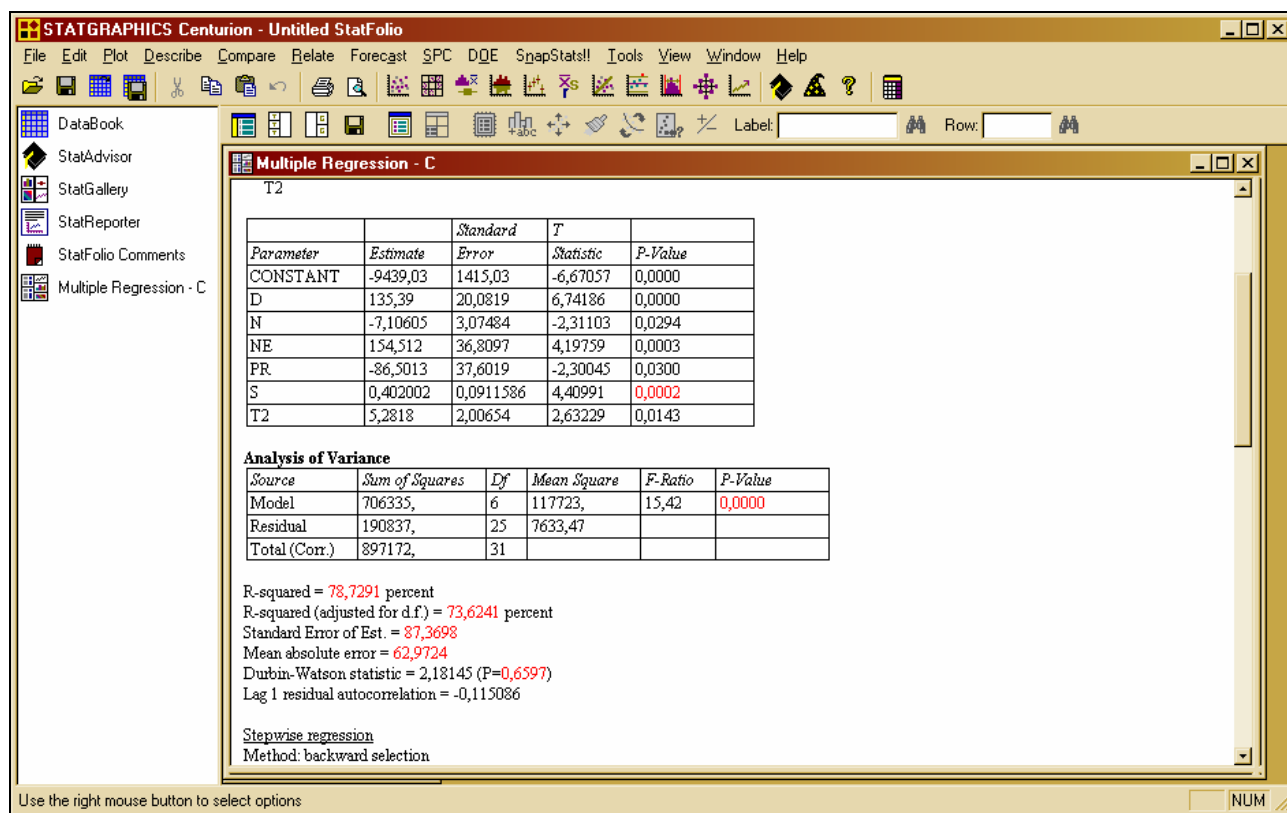
*StatAdvisor* сообщает, что независимая переменная **NE** (характеристика района) имеет большую значимость на 99 % достоверности. В соответствии со значением статистики R-квадрат указывает, что модель отражает 66,5817 % изменчивости переменной **C** ( $C = 6215,31 + 91,8416 \cdot D + 0,41636 \cdot S + 129,81 \cdot NE$ ). Стандартная ошибка равна 103,479. Средняя абсолютная ошибка, представляющая собой среднюю величину остатков, составляет 76,4689.

## ПОШАГОВЫЙ ОТБОР ПЕРЕМЕННЫХ

Оказывается, *Statgraphics* в состоянии сам отобрать наиболее значимые переменные. Удалим результаты анализа, которые вы получили. Снова получим первый анализ для всех переменных. Щелкните правой кнопкой мыши по сводке, выберите в меню *Analysis Options*, раскроется окно *Multiple Regression Options*.



В этом окне в разделе **Fit** выберите **Backward Selections**, нажмите ОК. Откроется новое окно анализа, результаты которого не очень хорошие –  $R$ -квадрат всего 73 %, зато все коэффициенты являются статистически значимыми.



Построенная модель отражает 96,074 % изменчивости переменной  $C$ .

## ЗАДАНИЯ

**Задача 1.** Для 12 стран приведены данные о валовом внутреннем продукте и расходах на образование (1980 г). Требуется найти функцию регрессии расходов на образование ( $EE$ ) на душу населения в зависимости от валового продукта ( $GDP$ ) на душу населения. Данные приведены в таблице. В столбце  $P$  приводится численность населения по каждой стране, столбцы  $EE/P$  и  $GDP/P$  заполните самостоятельно (подумайте, как это сделать рационально).

Страна	$EE$	$GDP$	$P$	$EE/P$	$GDP/P$
1	2	3	4	5	6
США	181,3	2586,4	227,64		
Япония	61,61	140,45	116,78		
Германия	38,62	815,00	61,56		

1	2	3	4	5	6
Франция	33,59	655,29	53,71		
Великобритания	29,9	534,97	55,95		
Италия	15,95	395,52	57,04		
Канада	18,9	261,41	23,94		
Бразилия	8,92	249,72	123,03		
Швеция	11,22	124,15	8,31		
Саудовская Аравия	6,4	115,97	8,37		
Финляндия	2,8	51,62	4,78		
Израиль	1,81	20,94	3,87		

Пусть в некоторой стране валовый продукт на душу населения вырос на 20 %. Спрогнозируйте рост удельных расходов на образование.

**Задача 2.** В файле *gbregr.sf* даны данные по Великобритании за  $n=20$  лет о потреблении цыплят ( $y$ ), среднедушевом доходе ( $x^{(1)}$ ), стоимости 1 фунта цыплят ( $x^{(2)}$ ), стоимости 1 фунта свинины ( $x^{(3)}$ ), стоимости 1 фунта говядины ( $x^{(4)}$ ). Требуется построить и сравнить уравнения регрессии вида:

1.  $y = a \cdot (x^{(2)})^b$ , функция спроса,
2.  $y = a \cdot (x^{(1)})^b$ , функция потребления,
3.  $y = a \cdot (x^{(1)})^b (x^{(2)})^c$ , функция спроса и потребления,
4.  $y = a \cdot (x^{(2)})^b (x^{(3)})^c (x^{(4)})^d$ , функция спроса с учетом цены на товарозаменители.

Следите за тем, насколько хорошо ваша модель приближает данные. Возможно, надо взять другие начальные значения ( $a$ ,  $b$ ).

Предположим, что в Великобритании цены на говядину возросли в 8 раз, а все остальные цены остались на прежнем уровне. Как изменится уровень потребления цыплят? Дайте прогноз, используя четвертую модель.