10. ПРОГРАММА АНАЛИЗА ДАННЫХ STATISTICA

10.1. Краткое описание системы STATISTICA

STATISTICA представляет собой интегрированную систему статистического анализа и обработки данных.

STATISTICA работает с четырьмя различными типами документов, которые соответствуют основным структурным компонентам системы:

– электронная таблица *Spreadsheet,* которая предназначена для ввода исходных данных и их преобразования;

– электронная таблица *Scrollsheet* для вывода численных и текстовых результатов анализа;

– график – документ в специальном графическом формате для визуализации и графического представления численной информации;

– отчет – документ в формате RTF *(Расширенный текстовой формат)* для вывода текстовой и графической информации.

☑ В соответствии со стандартами среды *Windows* каждый тип документа выводится в своем собственном окне в рабочей области системы *STATISTICA*. Как только это окно становится активным, изменяется панель инструментов и меню. В них появляются команды и кнопки, доступные для активного документа.

Запуск STATISTICA

Пуск ⇒ Программы ⇒ STATISTICA (рис. 10.1) ⇒

а) **Basic Statistics and Tables** (Основные статистики и таблицы) – запускается модуль системы Basic Statistics and Tables;

б) **STATISTICA** – появится переключатель модулей системы (рис. 10.2), в котором двойным щелчком левой кнопки мыши можно запустить нужный модуль, например тот же Basic Statistics and Tables.

В результате появится **Рабочее Окно** системы STATISTICA с меню соответствующего модуля (рис. 10.3), которое после щелчка мыши на свободном месте рабочего поля окна сворачивается в правый нижний угол (рис. 10.4).

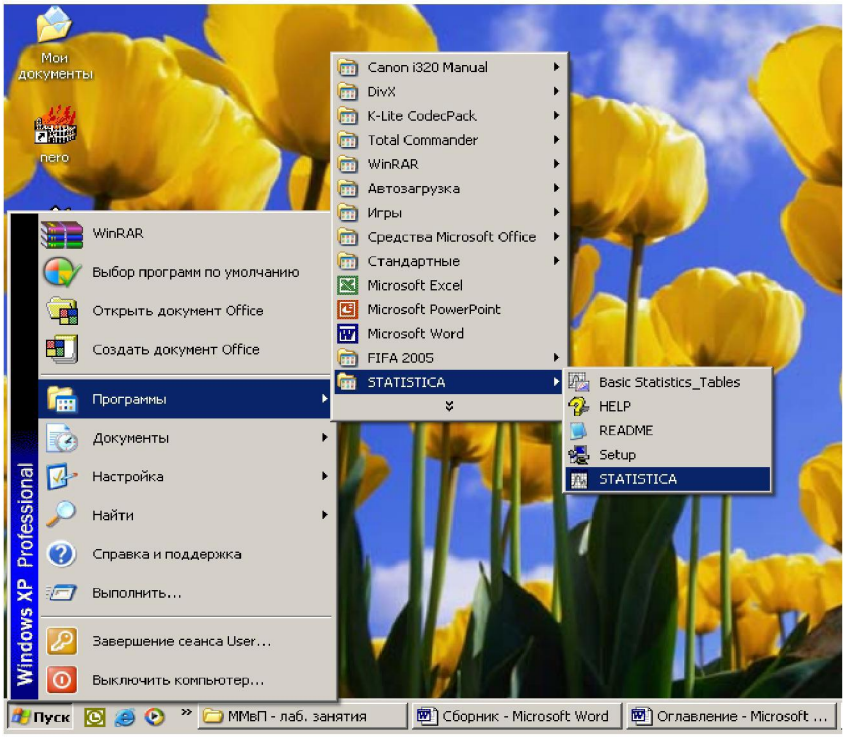


Рис. 10.1. STATISTICA 5.0 в меню Пуск

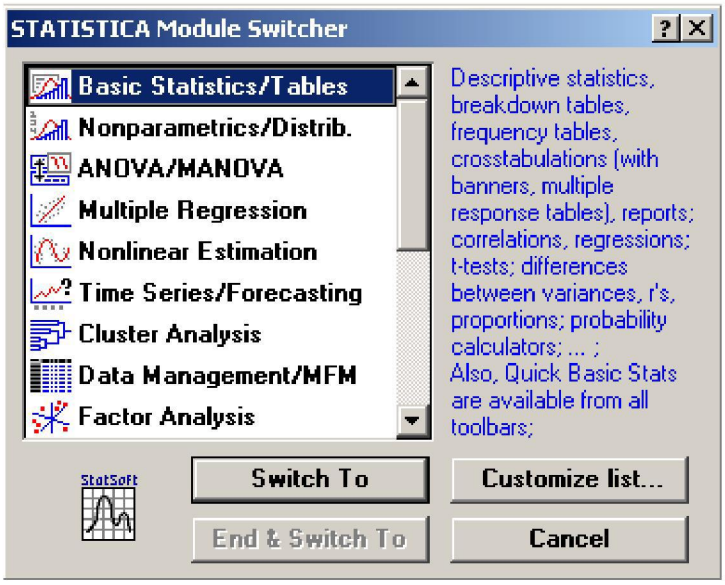


Рис. 10.2. Переключатель модулей STATISTICA 5.0

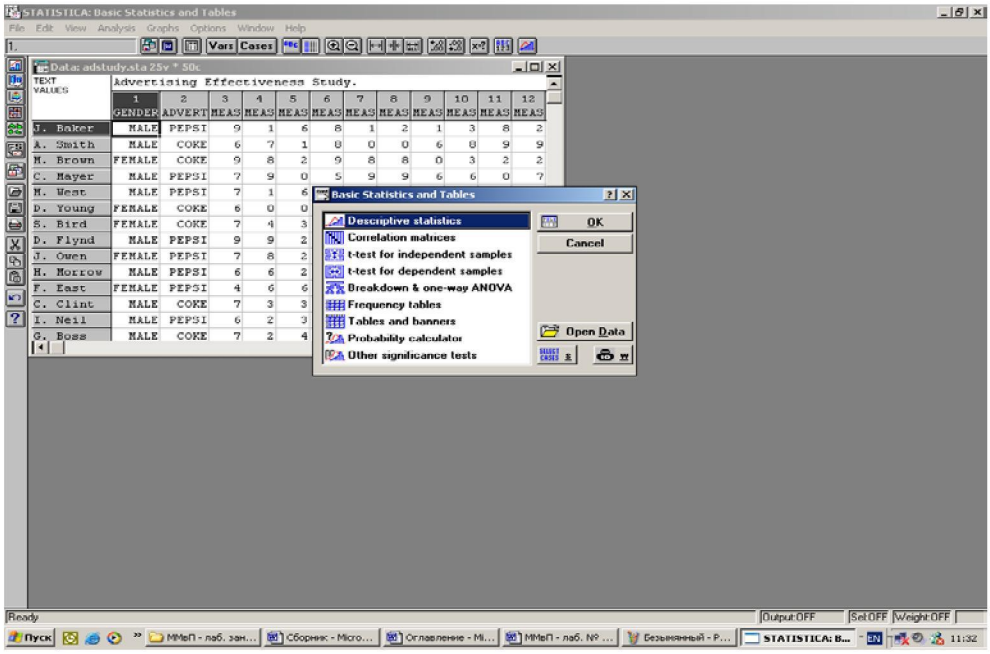


Рис. 10.3. Рабочее окно системы STATISTICA 5.0

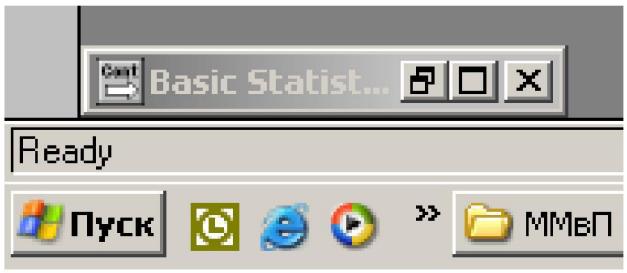


Рис. 10.4. Свернутое окно модуля
Basic Statistics and Tables

☑ При первом запуске STATISTICA (по умолчанию) автоматически открывается стандартный файл с данными *adstudy.sta*, который входит в набор примеров, поставляемых с системой (эти файлы находятся в каталоге stat\examples). При следующих запусках автоматически открывается последний файл, с которым вы работали в системе.

☑ В рабочей области может находиться только один файл с исходными данными и неограниченное число файлов с промежуточными результатами и графиками.

Исходные данные в системе STATISTICA организованы в виде электронной таблицы. Таблицы с исходными данными в STATISTICA носят особое название *SPREADSHEETS* и имеют расширение *.sta.

Электронная таблица системы *SPREADSHEETS* состоит из строк и столбцов. В отличие от обычных электронных таблиц, где строки и столбцы равноправны, в STATISTICA они имеют разные смысловые значения.

Столбцы электронной таблицы с исходными данными называются ***Variables (Переменные)***, а строки – ***Cases (Наблюдения)***. В качестве переменных обычно выступают исследуемые признаки (величины), а наблюдения – это значения, которые принимают переменные в отдельных измерениях.

Система может работать как с численными, так и с текстовыми данными. Аналогично *MS Excel* они поддерживают различные типы операций с данными, такие как операции с использованием *буфера обмена Windows*; операции с выделенными блоками значений, в том числе и с использованием метода *drag-and-drop*, автозаполнение блоков и т. д.

Рабочее окно имеет следующую структуру:

– верхний заголовок **STATISTICA: Basic Statistics and Tables** – запущен модуль Basic Statistics and Tables (Основные статистики и таблицы) (см. рис. 10.3). Если бы был запущен другой модуль, то его название указывалось бы в заголовке;

– строка меню;

– панель инструментов (под строкой меню и справа), соответствующая активному окну в рабочей области. На рис. 10.3, например, панель инструментов соответствует активному (и единственному) в данный момент окну с файлом исходных данных;

– рабочая область, занимающая большую часть окна, в которой выводятся все документы системы. На рис. 10.3, в частности, кроме меню модуля **Basic Statistics and Tables**, открыто окно с заголовком: **Data: adstudy.sta 25v*50c** – файл исходных данных (Data) с именем *adstudy.sta*, имеющий 25 столбцов (25v (*Variables*) – 25 переменных) и 50 строк (50c *(Cases)* – 50 наблюдений).

Создание файла данных

☑ Закрыть открытый файл данных, нажав кнопку  в правом верхнем углу окна (ненужного) файла данных (рис. 10.5).

В окне системы STATISTICA останется только строка меню, расположенная в верхней части окна (рис. 10.6).

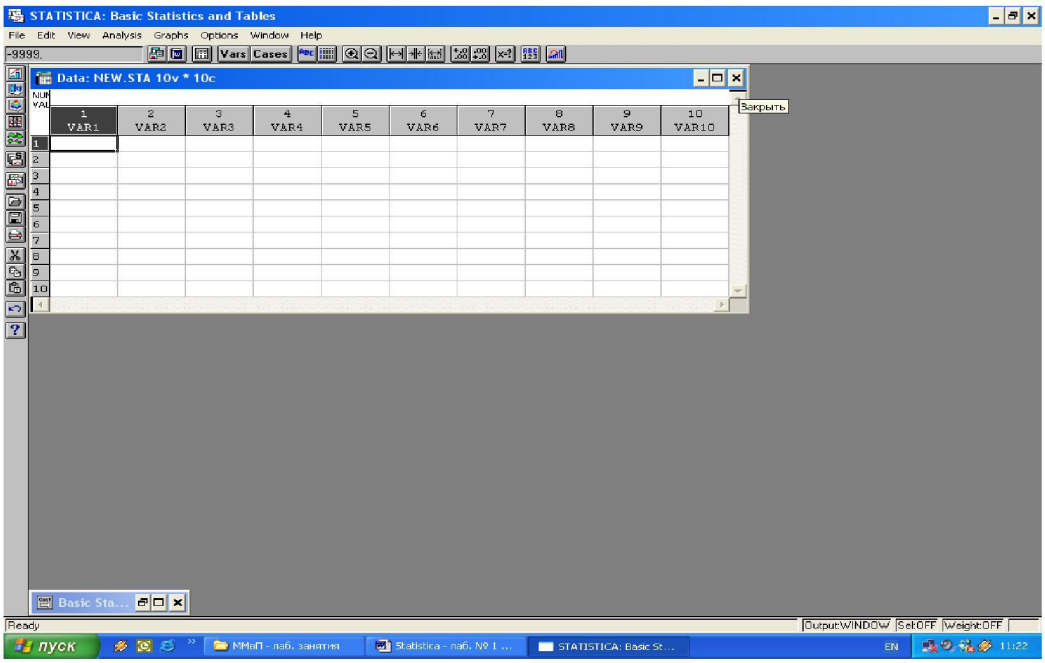


Рис. 10.5. Закрытие открытых файлов

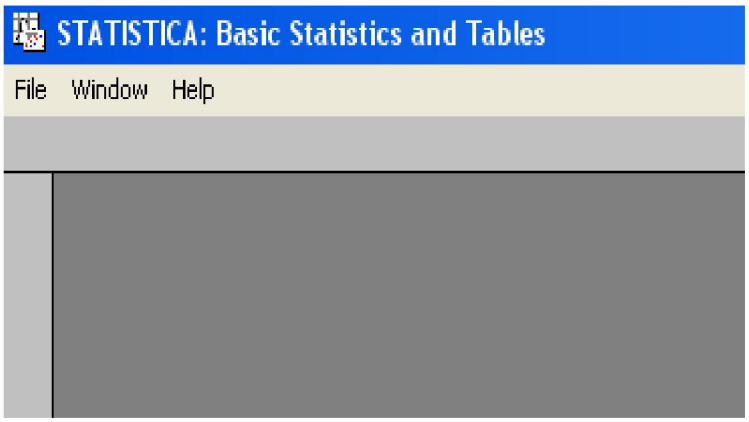


Рис. 10.6. Строка меню

Исходное положение: Вы находитесь в основном окне системы STATISTICA.

<u>Шаг 1.</u> Создание электронной таблицы.

В пункте меню **File (Файл)** выберите команду **New Data (Новые данные)** (рис. 10.7).

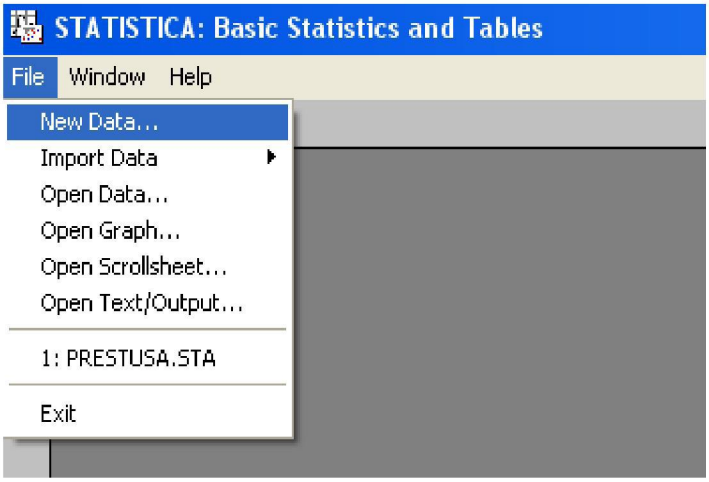


Рис. 10.7. Создание нового файла данных

В появившемся диалоговом окне **New Data: Specify File Name (Новые данные: Определить имя файла)** (рис. 10.8)

1) В поле **File Name (Имя файла)** введите имя нового файла (например, PRIMER1.STA)

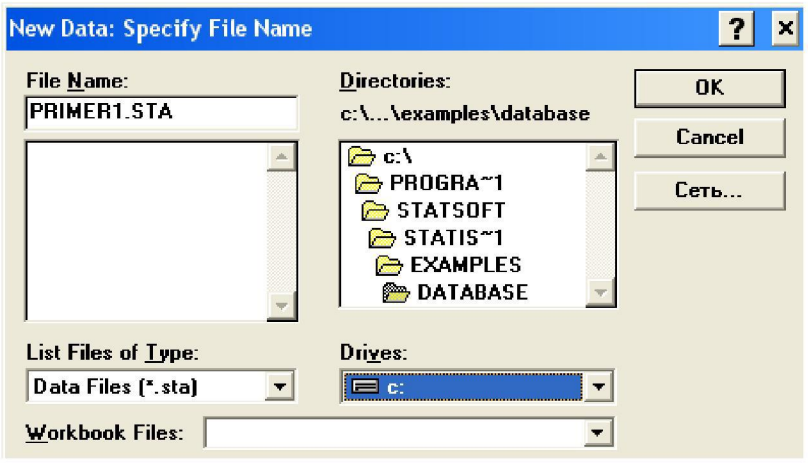


Рис. 10.8. Задание имени файла данных

☑ В системе STATISTICA 5.0 имена файлов (и переменных) задаются латинскими буквами и цифрами не более 8 символов.

2) В поле **Drives: (Драйвер (диск))** укажите логическое имя диска для сохранения файла данных: **Z:.**

3) В поле **Directories (Директория)** укажите папку для сохранения файла данных (например, **Z:\ММвП**).

4) Нажмите кнопку **OK** в правом углу окна. STATISTICA автоматически откроет пустую электронную таблицу с именем PRIMER1.STA (рис. 10.9).

В заголовке окна электронной таблицы автоматически отображается имя файла и его размер (PRIMER1.STA 10v * 10 c).

Размер таблицы по умолчанию принят 10 на 10 (10 переменных с именами VAR1, VAR2, …, VAR10 и 10 пронумерованных наблюдений).

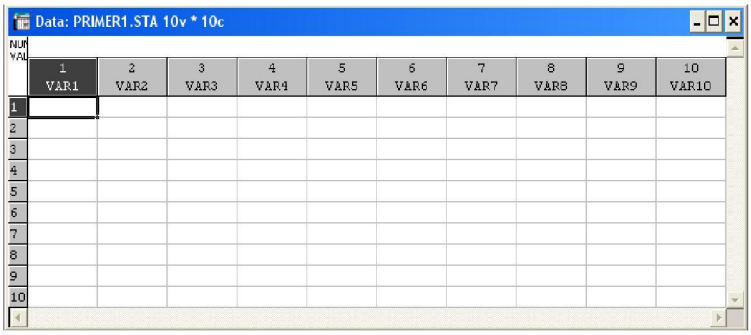


Рис. 10.9. Пустая электронная таблица для ввода данных

<u>Шаг 2.</u> Настройка размеров электронной таблицы.

Создадим столько переменных и наблюдений, сколько необходимо.

Для нашего примера требуются два переменных: *Абстрактное мышление* и *Вербальное мышление* и 40 наблюдений.

1) Нажмите кнопку Vars **Variables (Переменные)** на панели инструментов и выберите команду **Delete (Удалить).**

В диалоговом окне **Delete Variables (Удаление переменных)** укажите диапазон удаляемых переменных (**From variable (От переменной) – To variable (До переменной))**, как показано на рис. 10.10.

Нажмите кнопку **OK**.

☑ Чтобы упростить эту операцию, можно предварительно выделить переменные (столбцы), которые необходимо удалить.

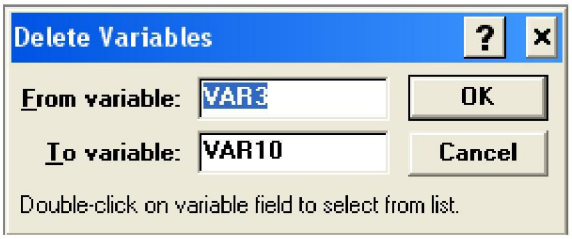


Рис. 10.10. Окно удаления лишних переменных (столбцов)

2) Нажмите кнопку Cases **Cases (Наблюдения)** на панели инструментов и выберите команду **Add (Добавить).**

В появившемся диалоговом окне **Add Cases (Добавление наблюдений)** укажите:

– количество добавляемых наблюдений (строк): **Number of Cases to Add (Количество наблюдений для добавления) – 40**;

– номер наблюдения, после которого вставить: **Insert after Cases (Вставить после наблюдения) – 10**, как показано на рис. 10.11.

Нажмите кнопку **OK**.

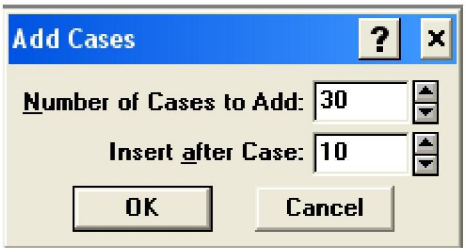


Рис. 10.11. Окно добавления наблюдений (строк) в таблицу

<u>Шаг 3.</u> Подготовка таблицы к вводу данных.

Заголовок таблицы.

После двойного щелчка на белом поле в таблице под словами: Data: PRIMER1.STA 2v * 40c на экране появится окно **Data File Header (Заголовок файла данных)**, в котором можно задать **заголовок таблицы (One-line Data File Header)** и дополнительную информацию о данных **(Data File Information/Notes)** (рис. 10.12).

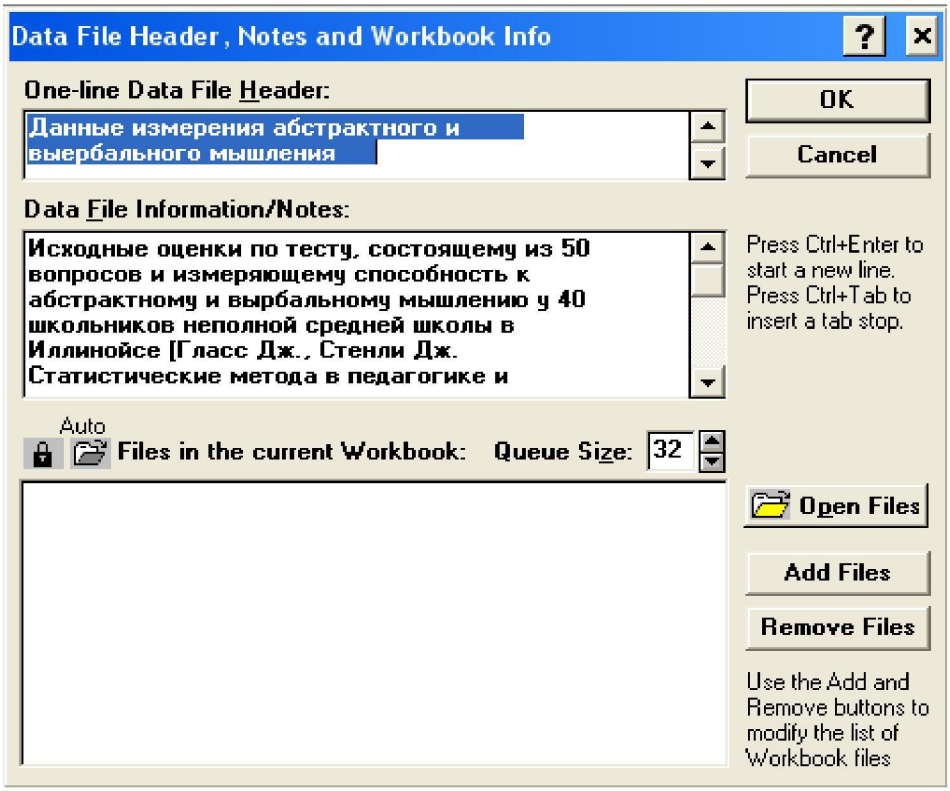
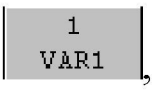


Рис. 10.12. Окно оформления заголовка таблицы

Имена переменных.

Для оформления имен и других спецификаций переменных можно:

а) дважды щелкнув на заголовке переменной: 

, задать спецификации переменных – каждой в отдельности (рис. 10.13):

Name (Имя): АБСТР (вместо VAR1);

Category (Тип): Number (Число);

Display Format (Формат отображения): 5 значащих цифр **(Column width)** и 0 десятичных знаков после запятой **(Decimals)** (вместо 8 и 3).

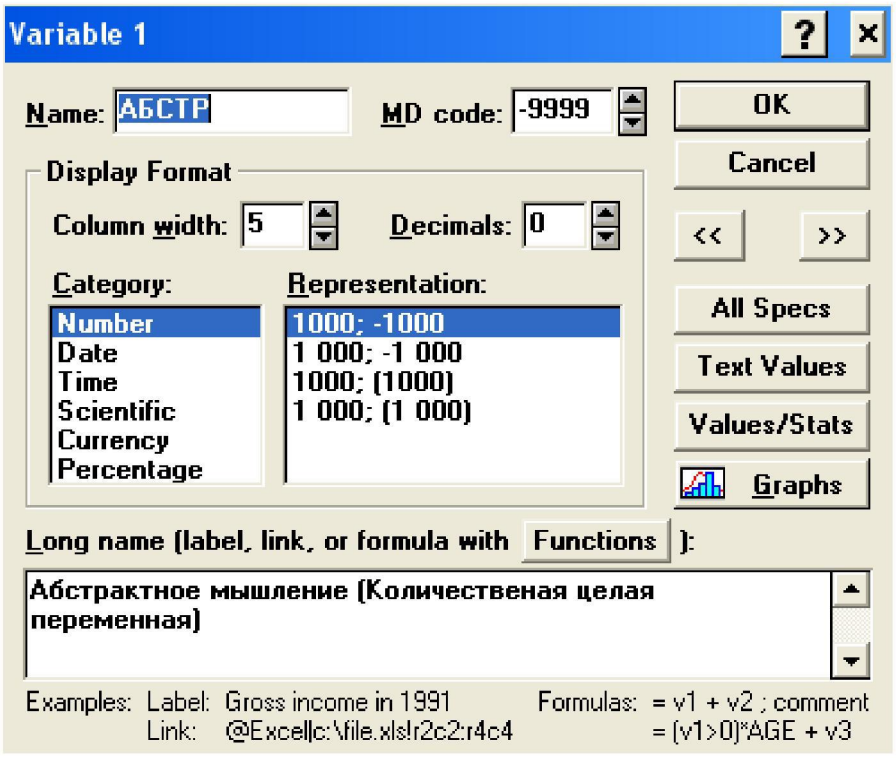


Рис. 10.13. Диалоговое окно **Спецификации переменной** (имя, тип, формат и т. д.)

б) нажав на панели инструментов кнопку **Specs of All Variable (Table) (Спецификации всех переменных (таблицы)),** задать задания спецификации всех переменных таблицы одновременно (рис. 10.14).

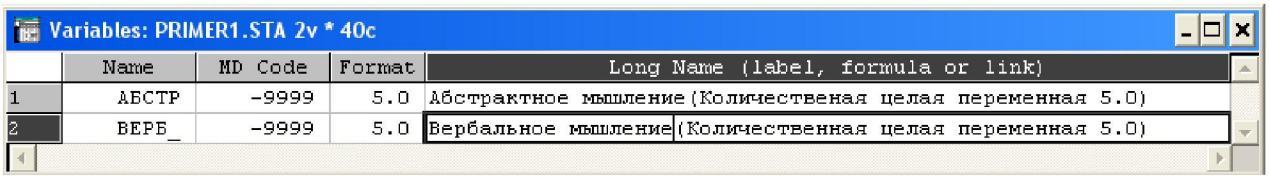


Рис. 10.14. Спецификации переменных

Имена наблюдений.

Нажмите кнопку Cases **Cases (Наблюдения)** на панели инструментов и выберите команду **Names (Имена).**

При первом выборе данного пункта появится диалоговое окно **Case Name Maneger (Менеджер имен случаев)** с запросом длины имен **(No case names in this file. Create? Width: 10. (Нет имен наблюдений в этом файле. Создать? Размер: 10)** (рис. 10.15). Введите подходящий размер поля имен наблюдений и нажмите кнопу **Yes (Да).**

В появившемся диалоговом окне **Case Name Maneger (Добавление наблюдений)** (рис. 10.16) введите имена наблюдений (респондентов). Нажмите кнопку **OK**.

Для того, чтобы имена наблюдений отображались в окне таблицы данных, необходимо нажать на панели инструментов кнопку – **Display Case Names on/off (Отображение имен наблюдений (вкл./выкл.))** (рис. 10.17).

Рис. 10.15. Вид окна

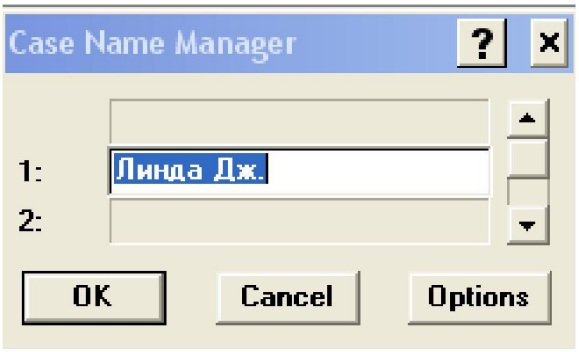


Рис. 10.16. Ввод имен наблюдений

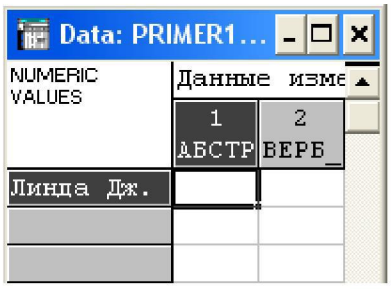


Рис. 10.17. Отображение таблицы

Шаг 4. Ввод данных в электронную таблицу.

Исходные данные наблюдений вводятся в таблицу с клавиатуры.

Шаг 5. Сохранение файла данных.

Для сохранения всех изменений и данных в таблице нажмите кнопку  – **Save Data File (Сохранить файл данных)** на панели инструментов, расположенной справа окна системы STATISTICA.

10.2. Проведение регрессионного анализа при помощи модуля Multiple Regressions

В стартовом диалоговом окне этого модуля (рис. 10.18.) при помощи кнопки **Variables** указываются зависимая (dependent) и независимые (ая) (independent) переменные. В поле **Input file** указывается тип файла с данными:

Raw Date – данные в виде строчной таблицы;

Correlation Matrix – данные в виде корреляционной матрицы.

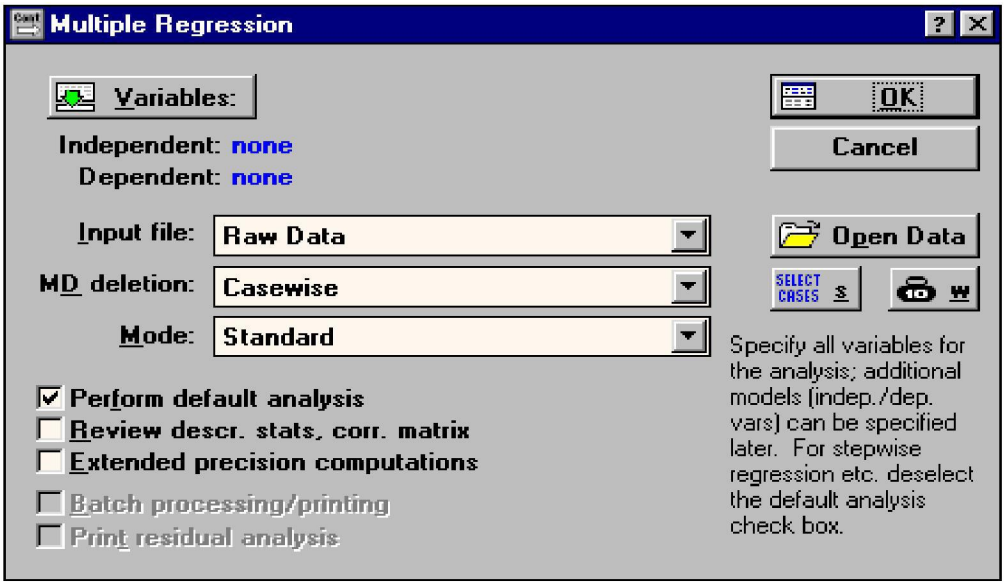


Рис. 10.18 . Стартовое диалоговое окно модуля Multiple Regressions

В поле **MD deletion** указывается способ исключения из обработки недостающих данных:

casewise – игнорируется вся строка, в которой есть хотя бы одно пропущенное значение;

mean Substitution – взамен пропущенных данных подставляются средние значения переменных;

pairwise – попарное исключение данных с пропусками из тех переменных, корреляция которых вычисляется.

В поле **Mode** указывается тип регрессионной модели:

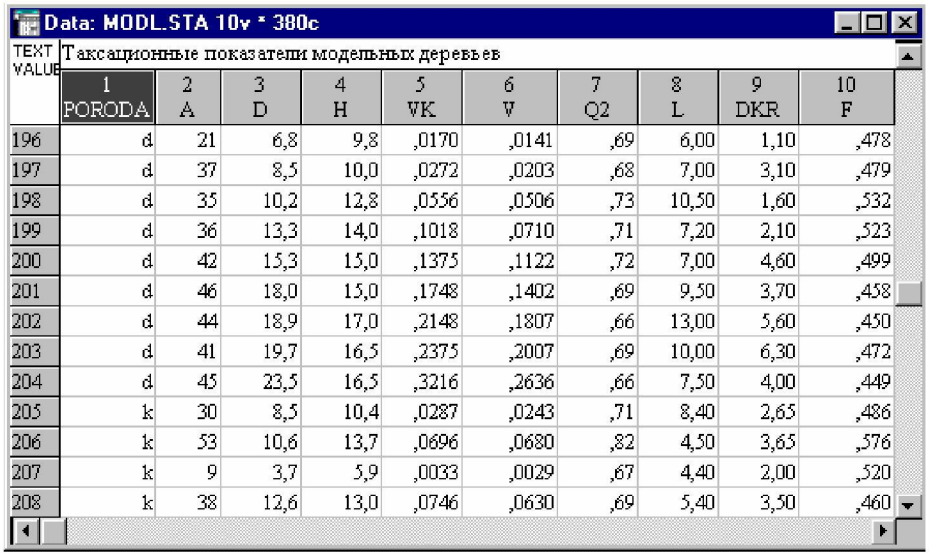
Standard – стандартная линейная модель вида

$Y = a_1 + a_2X_1 + a_3X_2 + \ldots + a_3X_3 + \ldots + a_nXn$

Fixed non linear – фиксированная нелинейная, т.е . нелинейная модель, но которая может быть приведена к линейному виду путем преобразования переменных.

Рассмотрим проведение регрессионного анализа на примере. Имеются данные обмера и таксации 380 модельных деревьев различных древесных пород. В файле данных (рис. 10.19) 10 переменных:

1	POROD	Древесня порода (d – дуб, lp – липа, k – клен, o – осина)
2	A	Возраст дерева, лет
3	D	Таксационный диаметр ствола дерева в коре, см
4	H	Высота дерева, м
5	VK	Объем ствола в коре, куб. м
6	V	Объем ствола без коры, куб. м
7	Q2	Второй коэффициент формы
8	L	Длина кроны дерева, м
9	DKR	Диаметр кроны дерева, м
10	F	Старое видовое число



	1 PORODA	2 A	3 D	4 H	5 VK	6 V	7 Q2	8 L	9 DKR	10 F
196	d	21	6,8	9,8	,0170	,0141	,69	6,00	1,10	,478
197	d	37	8,5	10,0	,0272	,0203	,68	7,00	3,10	,479
198	d	35	10,2	12,8	,0556	,0506	,73	10,50	1,60	,532
199	d	36	13,3	14,0	,1018	,0710	,71	7,20	2,10	,523
200	d	42	15,3	15,0	,1375	,1122	,72	7,00	4,60	,499
201	d	46	18,0	15,0	,1748	,1402	,69	9,50	3,70	,458
202	d	44	18,9	17,0	,2148	,1807	,66	13,00	5,60	,450
203	d	41	19,7	16,5	,2375	,2007	,69	10,00	6,30	,472
204	d	45	23,5	16,5	,3216	,2636	,66	7,50	4,00	,449
205	k	30	8,5	10,4	,0287	,0243	,71	8,40	2,65	,486
206	k	53	10,6	13,7	,0696	,0680	,82	4,50	3,65	,576
207	k	9	3,7	5,9	,0033	,0029	,67	4,40	2,00	,520
208	k	38	12,6	13,0	,0746	,0630	,69	5,40	3,50	,460

Рис.10.19. Вид окна с файлом данных

Найдем параметры регрессионного уравнения линейной связи объема ствола дуба в коре (переменная VK) от диаметра (D) и высоты (H) ствола. Вид уравнения: $VK = a_1 + a_2D + a_3H$.

Выставим опции стартового окна регрессионного анализа :

Variables: зависимая (dependent) переменная – VK; независимые (inde- pendent) – D, H (рис. 10.20); **Input file – Raw Date** (данные файла в виде строчной таблицы); **MD deletion – pairwise; Mode - Standard.**

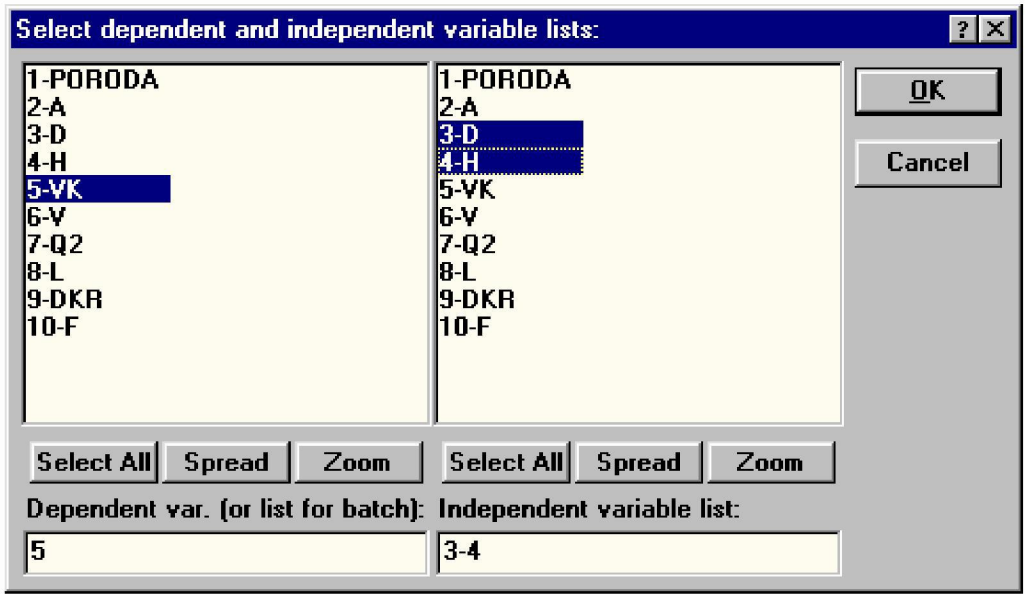


Рис. 10.20. Выбор зависимой и независимых переменных

Так как в файле данных содержится информация о модельных деревьях разных пород, а уравнение регрессии мы хотим получить для дуба, нужно воспользоваться кнопкой **Select cases** диалогового окна **Multiple Regressions**, чтобы установить условие включения случаев (строк файла данных) в статистическую обработку. В обработку должны включаться только те строки файла данных, для которых значение первой переменной V1 = 'd' (т. е. дуб) (рис. 10.21).

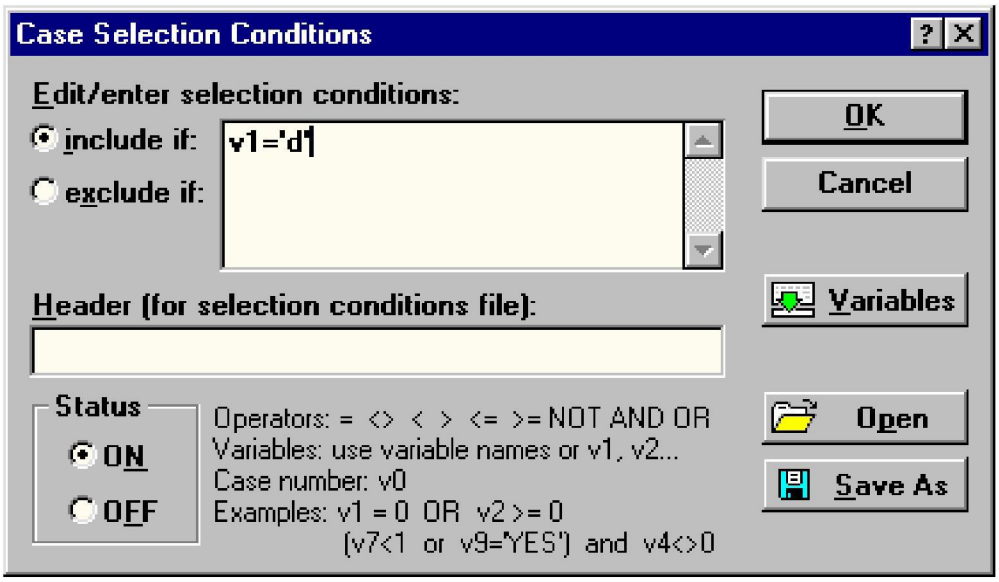


Рис. 10.21. Задание условия включения в обработку случаев со значением переменной V1 - дуб

После того, как все опции стартового диалогового окна регрессионного анализа выставлены, нажатие на кнопку **OK** приведет к появлению окна **Multiple Regressions Results** (результаты регрессионного анализа) (рис. 10.22), с помощью которого можно просмотреть результаты анализа в деталях.

В верхней части окна приводятся наиболее важные параметры полученной регрессионной модели:

Multiple R – коэффициент множественной корреляции (характеризует тесноту линейной связи между зависимой и всеми независимыми переменными. Может принимать значения от 0 до 1);

R^2или **RI** – коэффициент детерминации (численно выражает долю вариации зависимой переменной, объясненную с помощью регрессионного уравнения. Чем больше R^2, тем большую долю вариации объясняют переменные, включенные в модель);

adjusted R – скорректированный коэффициент множественной корреляции (этот коэффициент лишен недостатков коэффициента множественной корреляции. Включение новой переменной в регрессионное уравнение увеличивает RI не всегда, а только в том случае, когда частный F-критерий при проверке гипотезы о значимости включаемой переменной больше или равен 1. В противном случае включение новой переменной уменьшает значение RI и adjusted R^2);

adjusted R^2 или **adjusted RI** – скорректированный коэффициент детерминации (скорректированный R^2 можно с большим успехом (по сравнению с R^2) применять для выбора наилучшего подмножества независимых переменных в регрессионном уравнении);

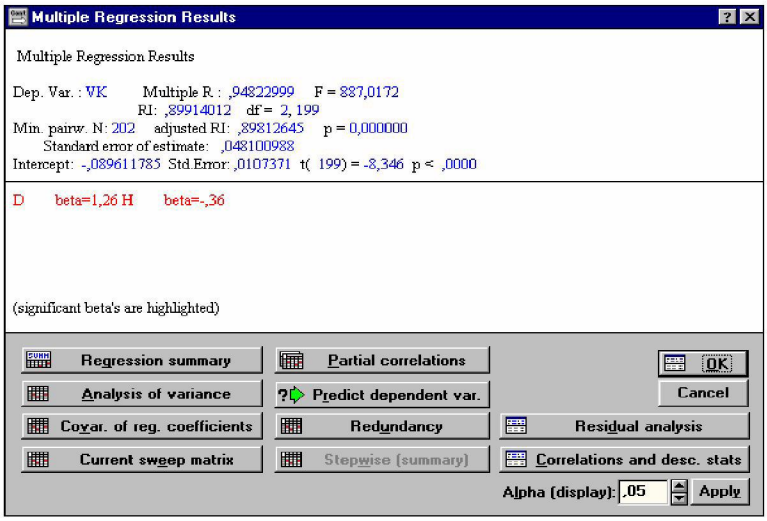


Рис. 10.22. Окно просмотра результатов регрессионного анализа

F – F-критерий;

df – число степеней свободы для F-критерия;

p – вероятность нулевой гипотезы для F-критерия;

Standard error of estimate – стандартная ошибка оценки (уравнения);

Intercept – свободный член уравнения;

Std.Error – стандартная ошибка свободного члена уравнения;

t – t-критерий для свободного члена уравнения;

p – вероятность нулевой гипотезы для свободного члена уравнения.

Beta – β-коэффициенты уравнения.

Это стандартизированные регрессионные коэффициенты, рассчитанные по стандартизированным значениям переменных. По их величине можно сравнить и оценить значимость зависимых переменных, так как β-коэффициент показывает на сколько единиц стандартного отклонения изменится зависимая переменная при изменении на одно стандартное отклонение независимой переменной при условии постоянства остальных независимых переменных. Свободный член в таком уравнении равен 0.

При помощи кнопок диалогового окна **Multiple Regressions Results** (рис. 10.22) результаты регрессионного анализа можно просмотреть более детально.

Кнопка **Regression summary** позволяет просмотреть основные результаты регрессионного анализа (рис. 10.23): **BETA** – β-коэффициенты уравнения; **St. Err. of BETA** – стандартные ошибки β-коэффициентов; **B** – коэффициенты уравнения регрессии; **St. Err. of B** – стандартные ошибки коэффициентов уравнения регрессии; **t (95)** – t-критерии для коэффициентов уравнения регрессии; **p-level** – вероятность нулевой гипотезы для коэффициентов уравнения регрессии.

Regression Summary for Dependent Variable: VK (modl.sta)

Continue...

R= ,94822999 RI= ,89914012 Adjusted RI= ,89812645
F(2,199)=887,02 p<0,0000 Std.Error of estimate: ,04810

N=202	BETA	St. Err of BETA	B	St. Err of B	t(199)	p-level	Valid N
Intercpt			-,090	,011	-8,35	,000	
D	1,257	,052	,027	,001	23,95	0,000	202,0
H	-,355	,052	-,012	,002	-6,77	,000	202,0

Рис. 10.23. Основные результаты регрессионного анализа

Таким образом, в результате проведенного регрессионного анализа получено следующее уравнение взаимосвязи между объемом ствола дуба в коре VK и диаметром D и высотой H ствола: VK = –0,090 + 0,027D – 0,012H. Все коэффициенты уравнения значимы на 5-процентном уровне (p-level < 0,05). Это уравнение объясняет 89,9 % (R^2= 0,899) вариации зависимой переменной. Ограничения модели: 2<=D>=31; 1,6<=H>=19,5.

Кнопка **Analysis of variance** позволяет ознакомиться с результатами дисперсионного анализа уравнения регрессии (рис. 10.24). В строках таблицы дисперсионного анализа уравнения регрессии – источники вариации: Regress. – обусловленная регрессией, Residual – остаточная, Total – общая. В столбцах таблицы: Sums of Squares – сумма квадратов, df – число степеней свободы, Mean Squares – средний квадрат, F – значение F-критерия, p-level – вероятность нулевой гипотезы для F-критерия.

F-критерий полученного уравнения регрессии значим на 5-процентном уровне. Вероятность нулевой гипотезы (p-level) значительно меньше 0,05, что говорит об общей значимости уравнения регрессии.

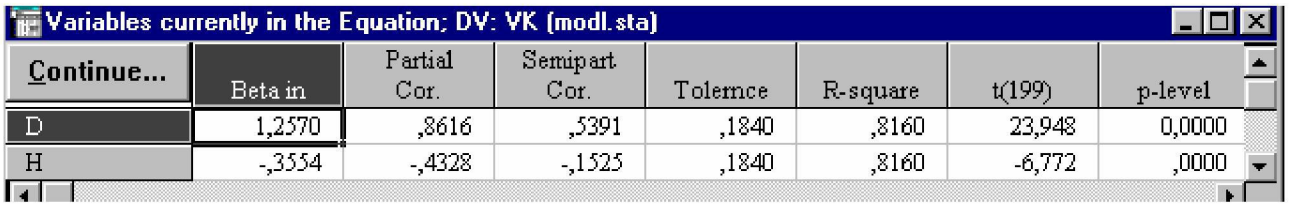
Analysis of Variance; DV: VK (modl.sta)

Continue...	Sums of Squares	df	Mean Squares	F	p-level
Regress.	4,104592	2	2,052296	887,0172	0,00
Residual	,460427	199	,002314		
Total	4,565019				

Рис. 10.24 .Результаты дисперсионного анализа уравнения регрессии

Кнопка **Partial correlations** позволяет просмотреть частные коэффициенты корреляции (Partial Cor.) между переменными (рис. 10.25). Частная корреляция – это корреляция между двумя переменными, когда

одна или больше из оставшихся переменных удерживаются на постоянном уровне (т. е. имеют постоянное значение). Частные коэффициенты корреляции, как и парные, могут принимать значения от –1 до +1.


Variables currently in the Equation; DV: VK (modl.sta)

Continue...	Beta in	Partial Cor.	Semipart Cor.	Tolernce	R-square	t(199)	p-level
D	1,2570	,8616	,5391	,1840	,8160	23,948	0,0000
H	-,3554	-,4328	-,1525	,1840	,8160	-6,772	,0000

Рис. 10.25. Результаты расчета частных коэффициентов корреляции

Сильная взаимная коррелированность независимых переменных в нашем уравнении затрудняет анализ влияния отдельных факторов на зависимую переменную. Отрицательный знак коэффициента уравнения перед высотой H, отрицательный знак частного коэффициента корреляции VK с H противоречат реальному положению дел. Положительный знак парного коэффициента корреляции между высотой и объемом ствола говорит о прямой взаимосвязи между ними.

В идеальной регрессионной модели независимые переменные вообще не коррелируют друг с другом. Однако в моделях, разрабатываемых для природных объектов, сильная коррелированность переменных является довольно частым явлением. Это приводит к увеличению ошибок уравнения, уменьшению точности оценивания, снижению эффективности использования регрессионной модели. Поэтому выбор независимых переменных, включаемых в регрессионную модель, должен быть очень тщательным.

Кнопка **Predict dependent var.** позволяет рассчитать по полученному регрессионному уравнению значение зависимой переменной по значениям независимых переменных. На рис. 10.26 приводится пример расчета объема ствола дуба в коре при диаметре ствола 14 см и высоте 11 м. Предсказанный (Predictd) объем составил 0,1614 куб. м.

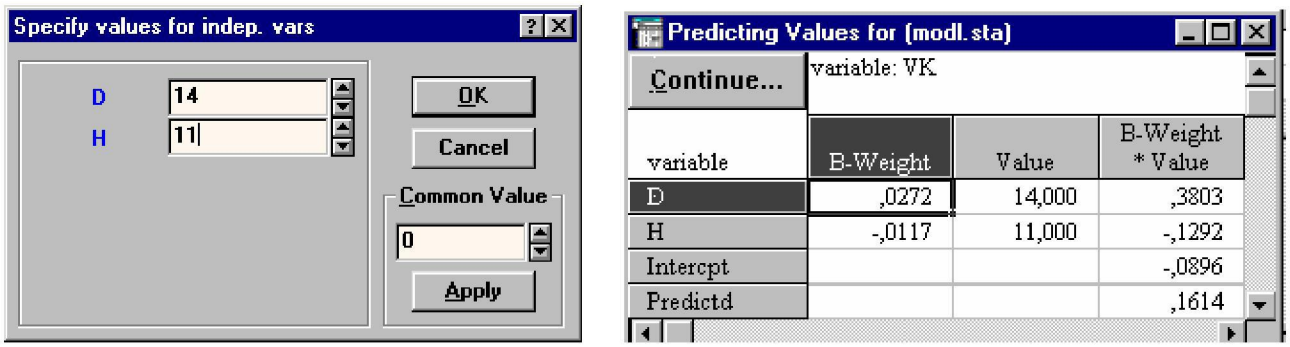


Рис. 10.26. Окно задания значений независимых переменных и результаты расчета по регрессионному уравнению зависимой переменной

Кнопка **Correlations and desc. stats** позволяет просмотреть описательные статистики и корреляционную матрицу с парными коэффициентами корреляции переменных, участвующих в регрессионной модели (рис. 10.27).

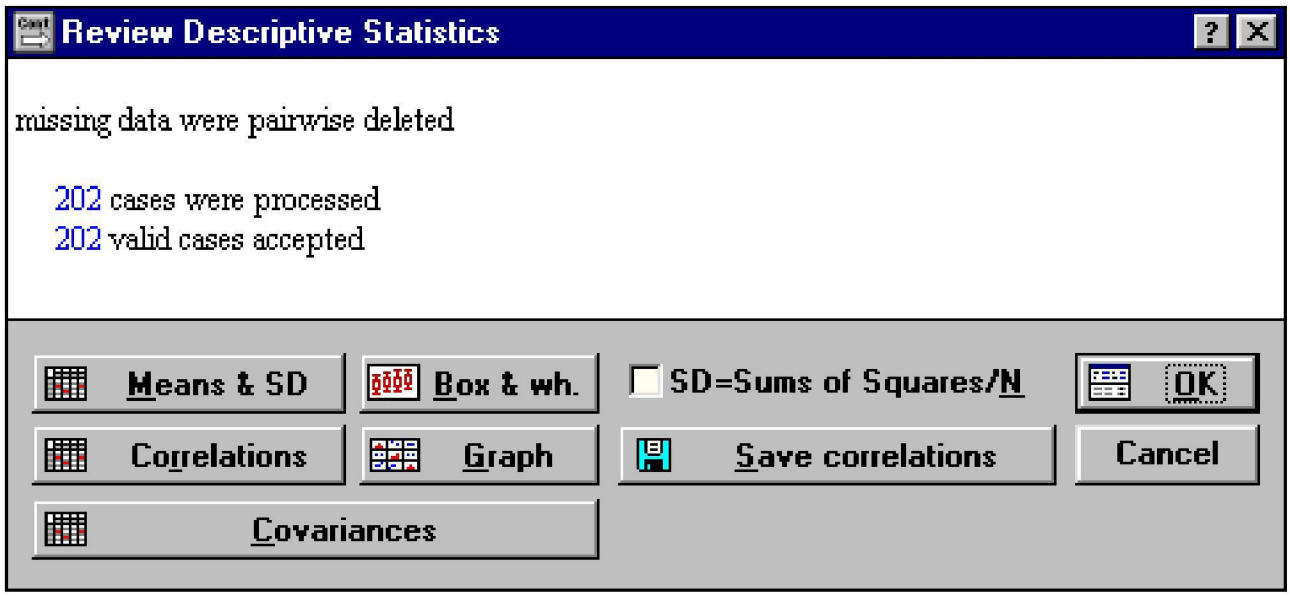


Рис. 10.27. Диалоговое окно Review Descriptive Statistics

Кнопка **Residual analysis** запускает процедуру всестороннего анализа остатков регрессионного уравнения (рис. 10.28). Остатки – это разности между опытными и предсказанными значениями зависимой переменной в построенной регрессионной модели.

Кнопка **Redundancy** предназначена для поиска выбросов. Выбросы – это остатки, которые значительно превосходят по абсолютной величине остальные. Выбросы показывают опытные данные, которые являются не типичными по отношению к остальным данным, и требуют выяснения причин их возникновения. Выбросы должны исключаться из обработки, если они вызваны ошибками регистрации, измерения. Для выделения имеющихся в регрессионных остатках выбросов предложен ряд показателей.

Показатель Кука (**Cook's Distance**) принимает только положительное значение и показывает расстояние между коэффициентами уравнения регрессии после исключения из обработки i-й точки данных. Большое значение показателя Кука указывает на сильно влияющий случай.

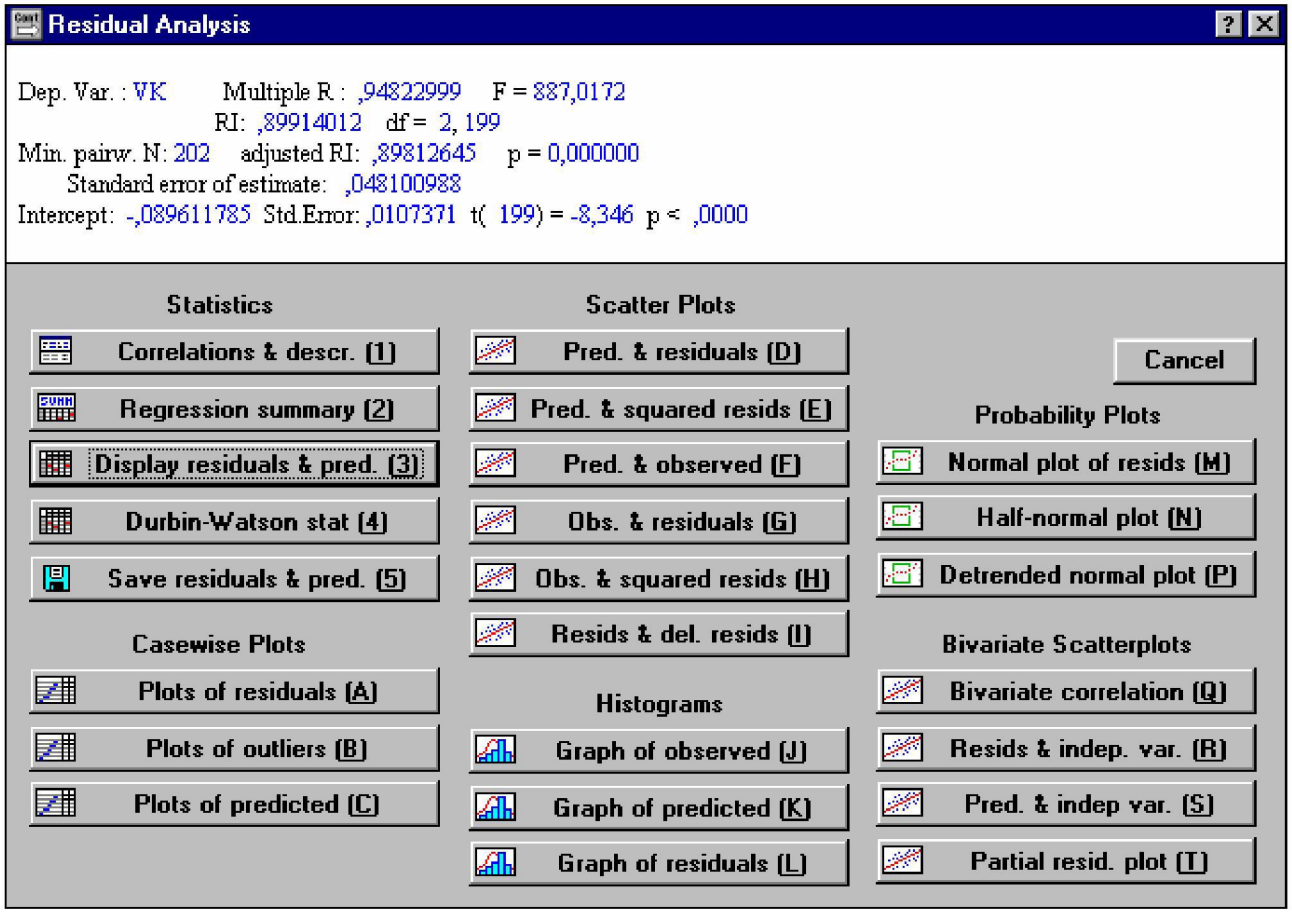


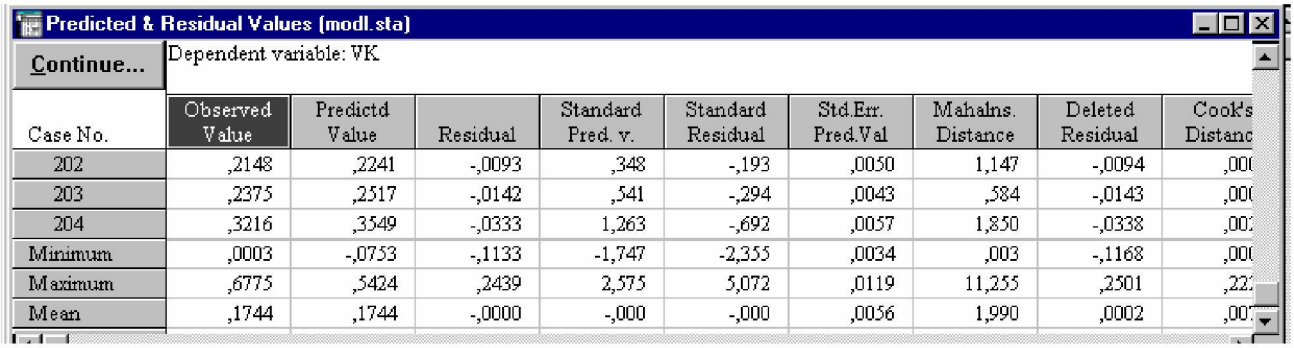
Рис. 10.28. Диалоговое окно Residual analysis (Анализ остатков)

Расстояние Махаланобиса (**Mahalns. Distance**) – показывает насколько каждый случай или точка в p-мерном пространстве независимых переменных отклоняется от центра статистической совокупности.

Внимательный анализ остатков позволяет оценить адекватность модели.

Остатки должны быть нормально распределены, со средним значением, равным нулю, и постоянной, независимы от величин зависимой и независимой переменных, дисперсий. Модель должна быть адекватна на всех отрезках интервала изменения зависимой переменной.

Просмотр величин остатков и специальных критериев, их оценивающих, осуществляется при помощи кнопки **Display residuals & pred.** окна Residual analysis. Для нашего примера фрагмент окна с этими данные представлен на рис. 10.29.



Case No.	Observed Value	Predictd Value	Residual	Standard Pred. v.	Standard Residual	Std.Err. Pred.Val	Mahalns. Distance	Deleted Residual	Cook's Distanc
202	,2148	,2241	-,0093	,348	-,193	,0050	1,147	-,0094	,000
203	,2375	,2517	-,0142	,541	-,294	,0043	,584	-,0143	,000
204	,3216	,3549	-,0333	1,263	-,692	,0057	1,850	-,0338	,00
Minimum	,0003	-,0753	-,1133	-1,747	-2,355	,0034	,003	-,1168	,000
Maximum	,6775	,5424	,2439	2,575	5,072	,0119	11,255	,2501	,22
Mean	,1744	,1744	-,0000	-,000	-,000	,0056	1,990	,0002	,00

Рис. 10.29. Окно со значениями остатков (Residuals), показателями Кука (Cook's Distance), расстоянием Махаланобиса (Mahalns. Distance), опытными (Observed Value) и предсказанными по уравнению (Predictd Value) значениями зависимой переменной

Вполне достаточно бывает одного графического анализа остатков. О нормальности остатков можно судить по графику остатков на нормальной вероятностной бумаге. Чем ближе распределение к нормальному виду, тем лучше значения остатков ложатся на прямую линию. График строится при помощи кнопки **Normal plot of resids.** окна Residual analysis (рис. 10.30).

Важно просмотреть графики зависимости остаток от каждой из независимых переменных. Их легко просмотреть при помощи кнопки **Resids & indep. Var.** окна Residual analysis. Остатки должны быть нормально распределены, т. е. на графике они должны представлять приблизительно горизонтальную полосу одинаковой ширины на всем ее протяжении. Коэффициент корреляции r между регрессионными остатками и переменными должен равняться нулю.

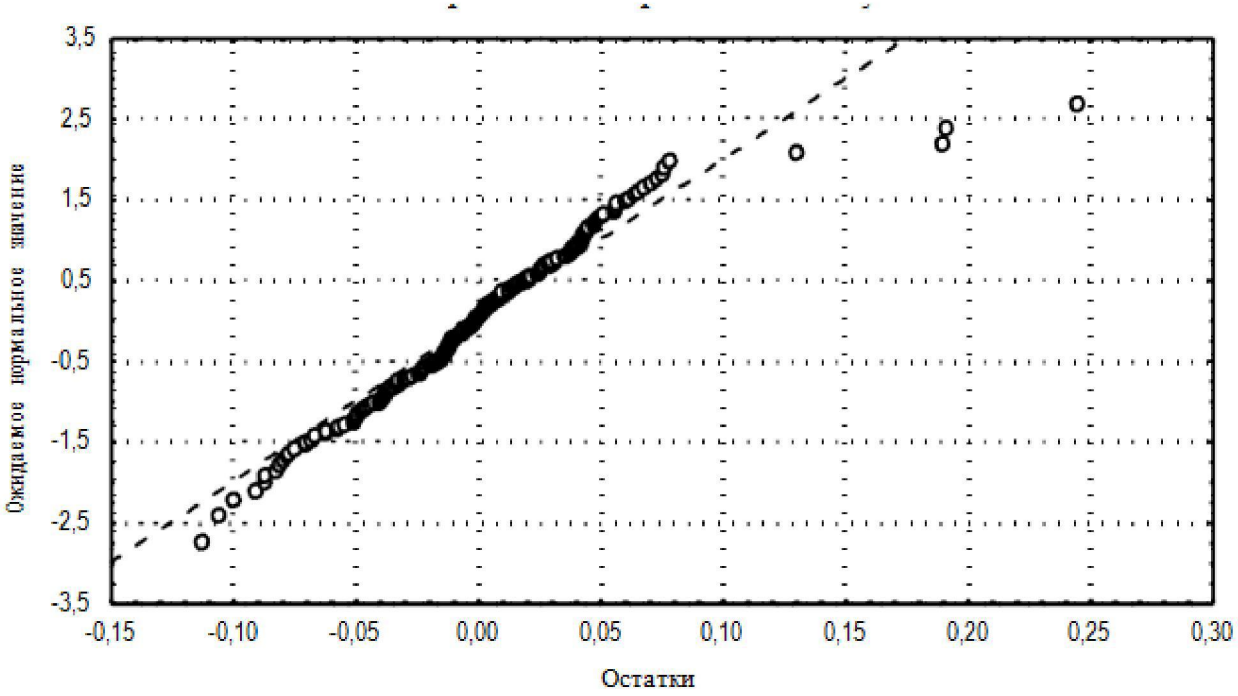


Рис. 10.30. График остатков на нормальной вероятностной бумаге

В нашем случае на графиках остатков (рис. 10.31) хорошо просматривается нелинейный тренд, что вызывает сомнение в адекватности модели. Присутствие нелинейного тренда в регрессионных остатках говорит о необходимости пересмотра модели (преобразования или ввода новых переменных, перехода от линейной модели к нелинейной).

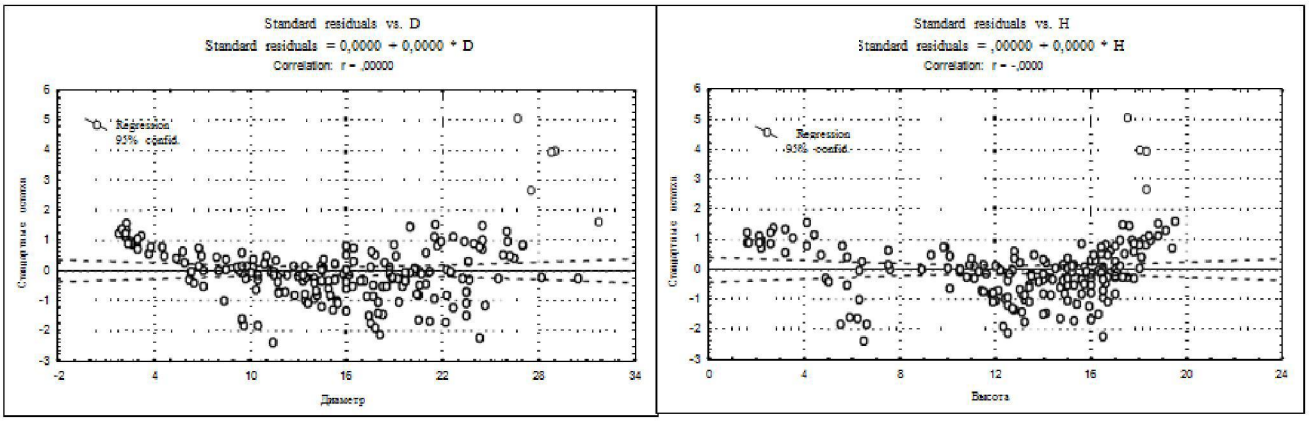


Рис. 10.31. Зависимость остатков от независимых переменных: диаметра и высоты

Для выявления нестабильности дисперсии ошибки уравнения при помощи кнопки **Pred. & residuals** окна Residual analysis можно создать график зависимости регрессионных остатков от предсказанного значения зависимой переменной. Рис. 10.32 позволяет сделать заключение о непостоянстве дисперсии ошибки уравнения (с увеличением значений зависимой переменной дисперсия увеличивается). Это еще одно подтверждение неадекватности анализируемой модели.

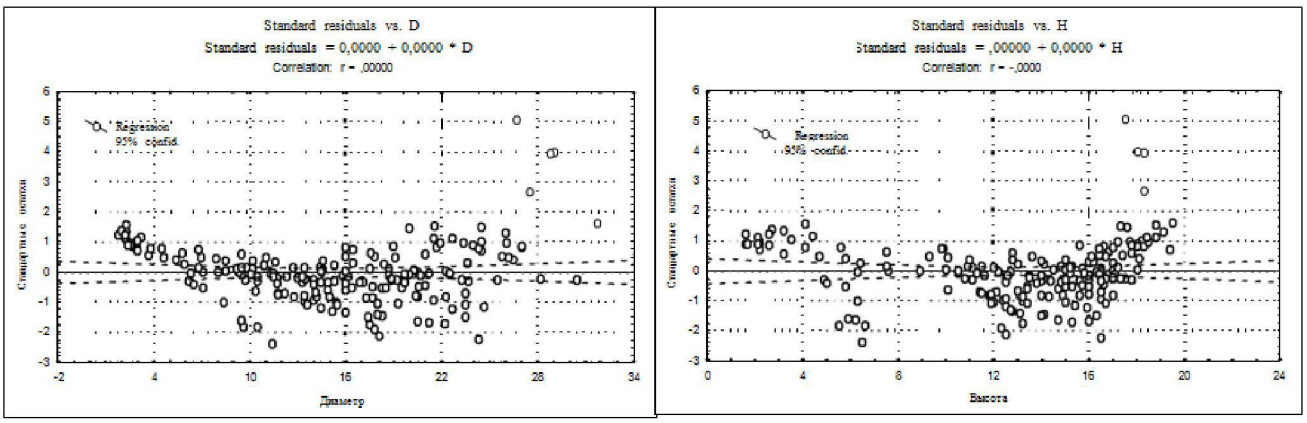


Рис. 10.32. Зависимость регрессионных остатков от предсказанных значений зависимой переменной

Очень удобным визуальным способом оценки адекватности регрессионной модели является анализ графического изображения опыт-

ных и полученных по регрессионному уравнению значений зависимой переменной. Изображение строится нажатием кнопки **Pred. & observed** окна Residual analysis.

Из рис. 10.33 хорошо видно, что линейный вид нашей модели плохо описывает взаимосвязь объема ствола дуба в коре от его диаметра и высоты (модель при малых и больших значениях отклика занижает величину зависимой переменной). Эта связь носит нелинейный характер.

Рассмотрим порядок нахождения коэффициентов уравнений регрессии нелинейного вида, которые через преобразования переменных могут быть приведены к линейной модели. Найдем параметры регрессионного уравнения связи объема ствола дуба в коре (переменная VK) от диаметра D ствола. Вид уравнения: $VK = a_1 + a_2D + a_3D^2$.

Опцию **Mode** стартового окна регрессионного анализа (см. рис. 10.18) выставим в положение **Fixed non linear**.

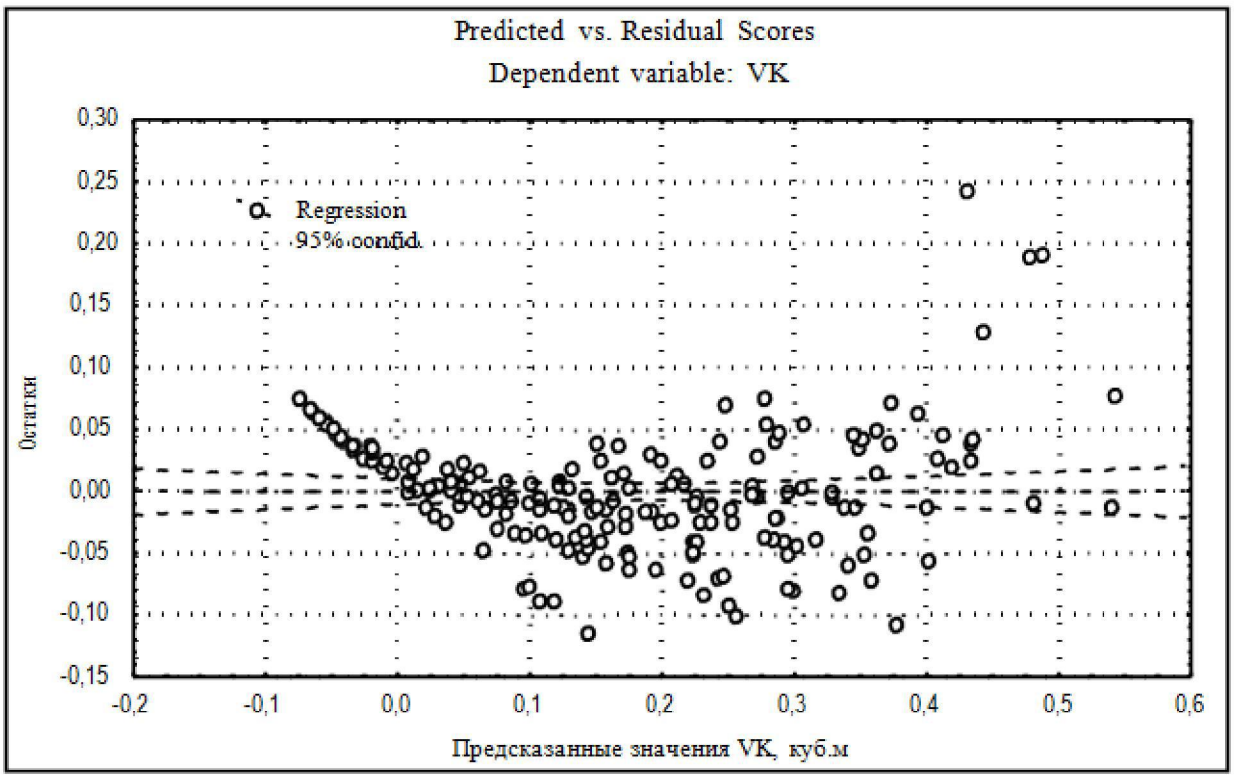


Рис. 10.33. Линия регрессии, опытные и полученные по регрессионному уравнению значений зависимой переменной

Если выбран фиксированный нелинейный тип регрессионной модели, то после нажатия на кнопку ОК в диалоговом окне **Multiple Regressions** (рис. 10.34) появляется окно **Non-linear Components Regression**, в котором можно выбрать следующие типы преобразования переменных: X^2, X^3, X^4, X^5, $\sqrt{X}$ ($X \geq 0$), $\ln X$ ($X > 0$), $\lg_{10} X$ ($X > 0$), e^X ($40 < X < -40$), 10^X, $1/X$ ($X \neq 0$). Если потребуются какие либо иные преобразования переменных, то тогда в файле данных следует создать мнимые вычисляемые переменные и включить их в качестве зависимых переменных в регрессионную модель.

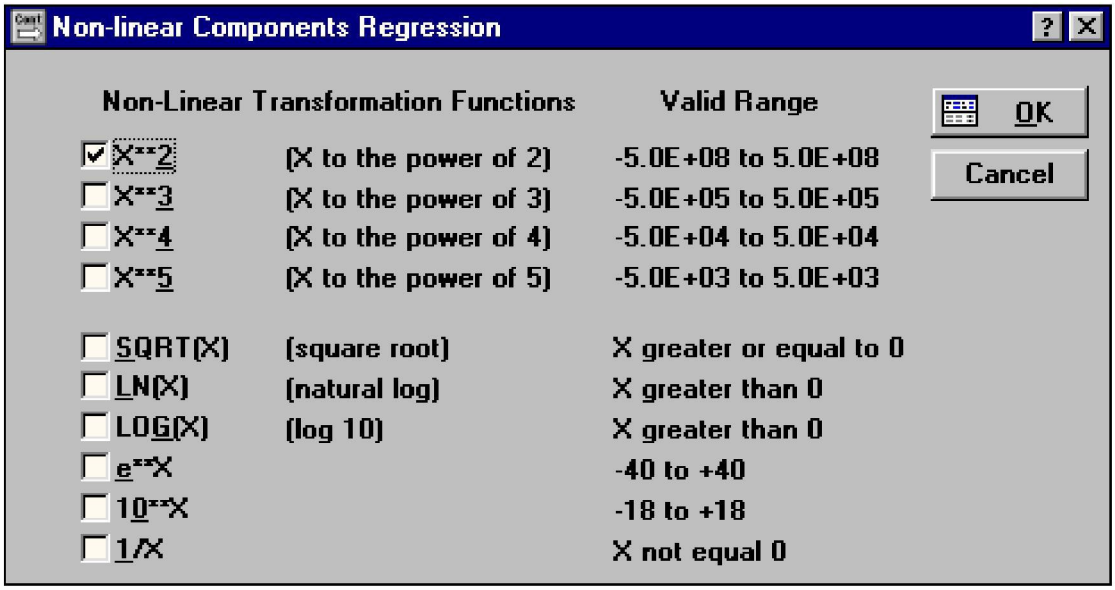


Рис. 10.34. Окно выбора типов преобразования переменных

После того, как тип преобразования переменных определен (в нашем примере это возведение в квадрат), необходимо уточнение зависимой и независимых переменных фиксированной нелинейной регрессионной модели. Оно проводится на следующем шаге при помощи кнопки **Variables** диалогового окна **Model Definition** (Уточнение модели) (рис. 10.35).

Зависимой (**dependent**) переменной в нашем случае будет VK; независимыми (**independent**) – D и D^2 (рис. 10.36). Переменная D^2 значит-

ся в списке переменных как V3**2, так как переменная D является третьей в списке переменных.

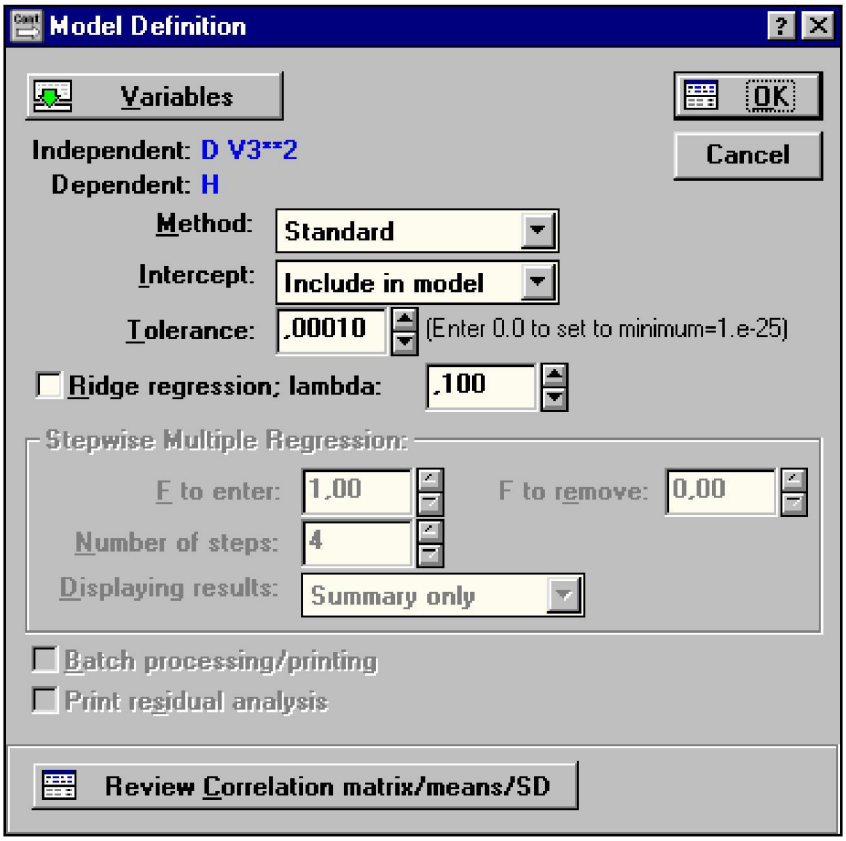


Рис. 10.35. Диалоговое окно Model Definition (Уточнение модели)

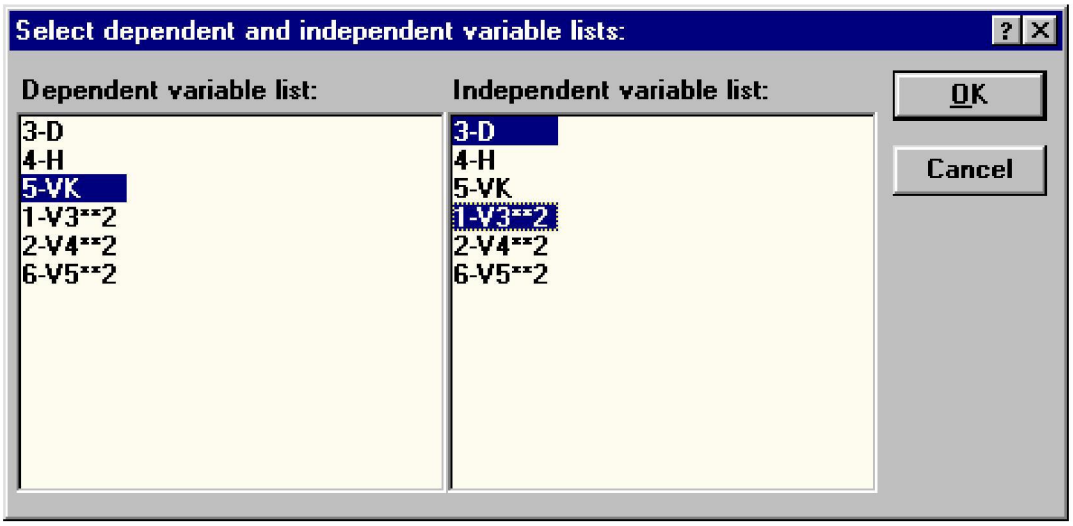


Рис. 10.36. Выбор переменных для расчета уравнения

$$VK = a_1 + a_2D + a_3D^2$$

Уравнение взаимосвязи между объемом ствола дуба в коре (VK) от его диаметром (D) оказалось следующее: VK = 0,00023 – 0,0034D + + 0,0008D^2. Все коэффициенты уравнения (за исключением свободного члена) значимы на 5-процентном уровне (p-level < 0,05). Это уравнение объясняет 95,8 % (R^2 = 0,958) вариации зависимой переменной (рис. 10.37).

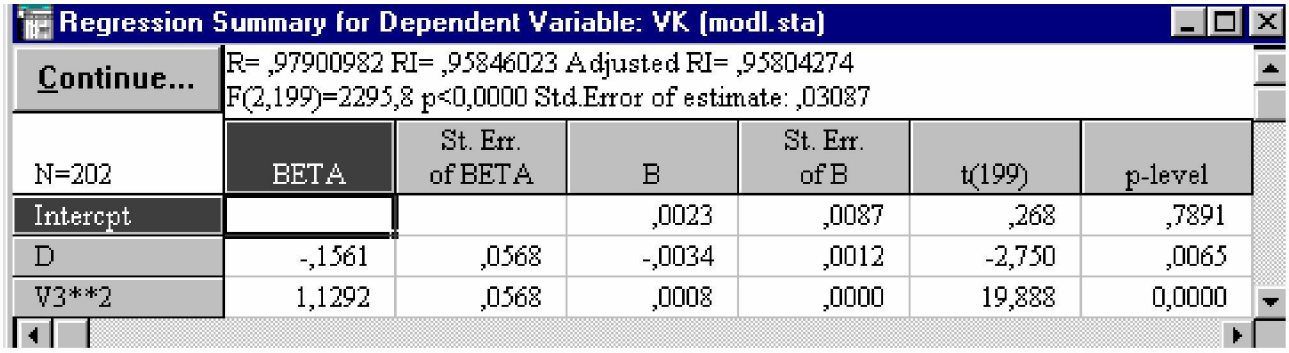


Рис. 10.37. Результаты регрессионного анализа модели $VK = a_1 + a_2D + a_3D^2$

По всем стандартным параметрам второе уравнение регрессии значительно лучше первого. Это наглядно подтверждает и график на рис. 10.38.

Найдем параметры еще одного регрессионного уравнения: $VK = a_1D^{a_2}H^{a_3}$. Это степенное уравнение может быть приведено к линейному виду через логарифмирование: $\ln VK = \ln a_1 + a_2 \ln D + a_3 \ln H$.

При помощи кнопки **Variables** укажем зависимую VK и независимые переменные – D, H. Опцию Mode стартового окна регрессионного анализа выставим в положение **Fixed non linear**. В качестве типа преобразования переменных выберем натуральный логарифм (ln (X)). В диалоговом окна **Model Definition** при помощи кнопки **Variables**

уточним модель, переопределив зависимую и независимые переменные так, как это показано на рис. 10.39.

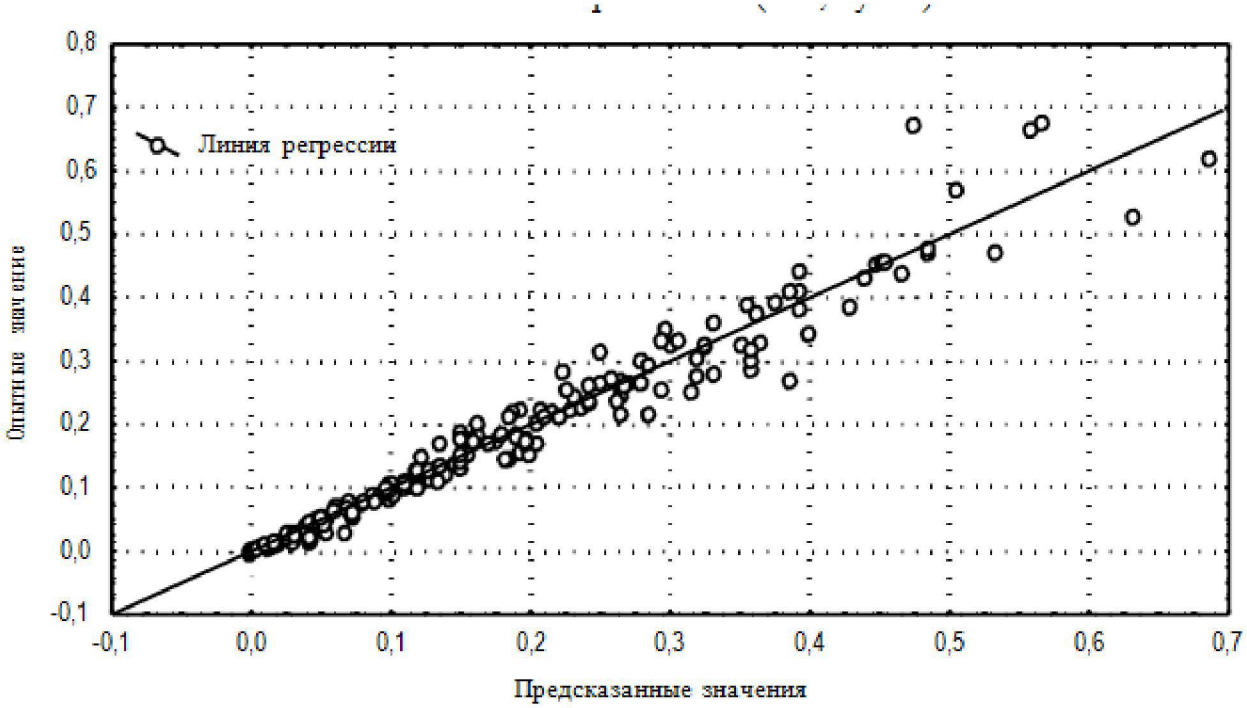


Рис. 10.38. Линия регрессии, опытные и полученные по регрессионному уравнению значений зависимой переменной

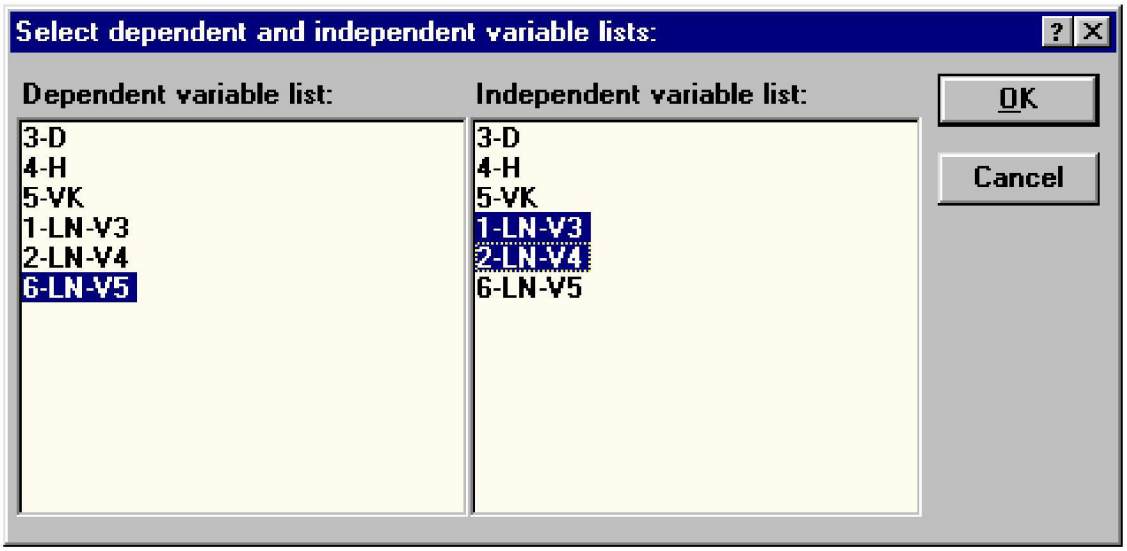


Рис. 10.39. Выбор переменных для расчета уравнения

$$\ln VK = \ln a_1 + a_2 \ln D + a_3 \ln H$$

Основные результаты регрессионного анализа представлены на рис. 10.40.

Regression Summary for Dependent Variable: LN-V5 (modl.sta)

Continue...

R= ,99788945 RI= ,99578335 Adjusted RI= ,99574098
F(2,199)=23497, p<0,0000 Std.Error of estimate: ,11405

N=202	BETA	St. Err. of BETA	B	St. Err. of B	t(199)	p-level
Intercpt			-9,87890	,036829	-268,233	0,00
LN-V3	,682935	,013811	1,87395	,037898	49,447	0,00
LN-V4	,327696	,013811	1,03462	,043606	23,726	0,00

Рис. 10.40. Результаты регрессионного анализа модели

$$\ln VK = \ln a_1 + a_2 \ln D + a_3 \ln H$$

Уравнение выглядит следующим образом: lnVK = -9,8789 + + 1,8739lnD + 1,0346lnH или в степенном виде: $VK = 0{,}00005\ D^{1{,}8739} \times \times H^{1{,}0346}$. Все коэффициенты уравнения значимы на 5-процентном уровне (p-level < 0,05). Это уравнение объясняет 99,6 % ($R^2 = 0{,}996$) вариации зависимой переменной. Ошибка уравнения 0,11405. Чтобы выразить ее в процентах, сравним абсолютную величину ошибки со средним значением зависимой переменной (lnVK): 0,11405/2,46166 · 100% = = 4,6%.

Проверим адекватность полученной модели через анализ остатков. В целом он даст положительное заключение. В качестве иллюстрации приведем лишь несколько графиков (рис. 10.41, 10.42), подтверждающих такой вывод.

Поиск наилучшей регрессионной модели представляет собой довольно громоздкий процесс. При помощи опции **Method** пользователь может отказаться от стандартного проведения регрессионного анализа (Standard) и воспользоваться методами пошагового включения переменных в регрессионную модель (**Forward stepwise**) или пошагового исключения переменных (**Backward stepwise**) из регресси-

онной модели. Опция **Displaying results** позволяет просматривать только итоговые результаты регрессионного анализа (Summary only) или после каждого шага включения или исключения переменных (At each step). Если необходимо получить регрессионную модель без свободного члена уравнения, тогда в списке поля **Intercept** нужно выбрать **Set to zero.**

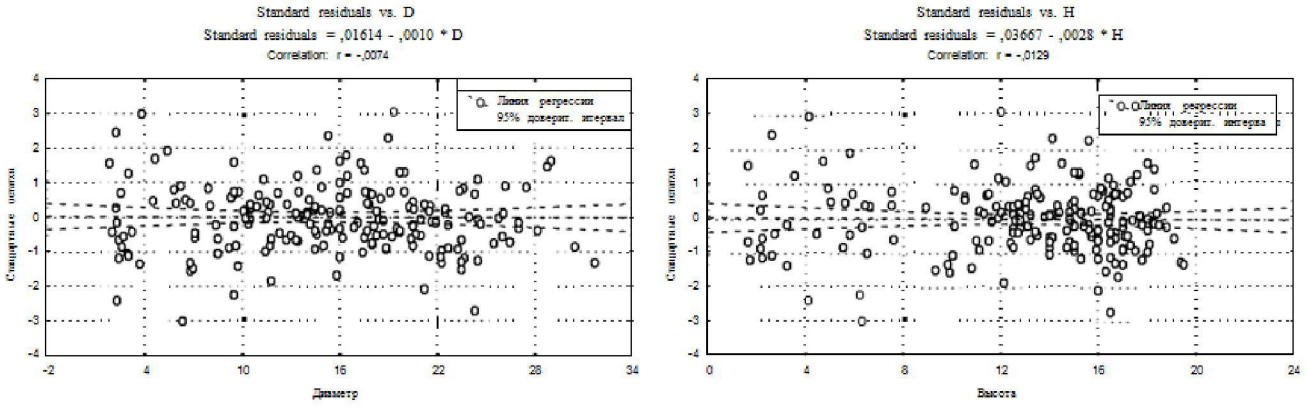


Рис. 10.41. Зависимость остатков степенного уравнения от независимых переменных: диаметра и высоты

Воспользуемся методом пошагового включения переменных для нахождения наилучшего регрессионного уравнения, описывающего объем ствола дуба в коре (VK). В качестве независимых переменных, которые потенциально могут быть включены в модель, примем: диаметр ствола D, квадрат диаметра D^2, высота ствола H, квадрат высоты ствола H^2, произведение диаметра ствола на его высоту DH, квадрат произведения диаметра ствола на его высоту DH^2.

Вначале создадим новую переменную DH. В файле данных она будет одиннадцатой по счету. Для расчета значений этой переменной вызовем окно с экспликацией этой переменной (рис. 10.43) и в поле **Long name** введем формулу, в соответствии с которой значения переменной должны быть рассчитаны, т. е =V3 · V4".

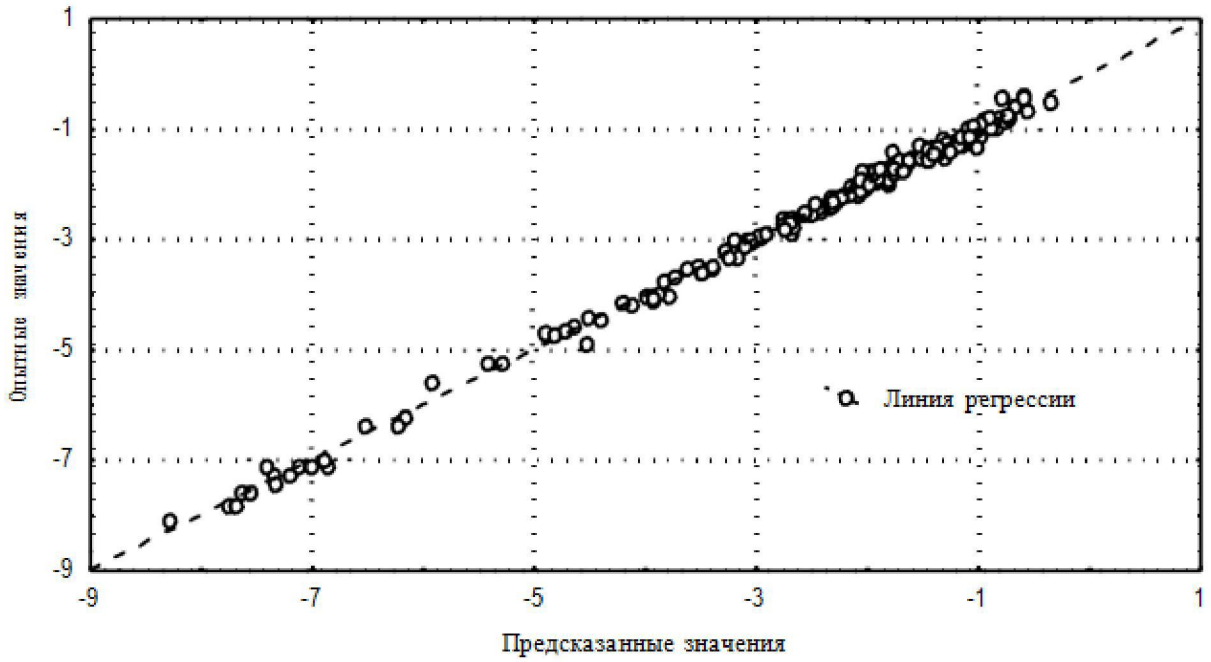


Рис. 10.42. Линия регрессии, опытные и полученные по степенному регрессионному уравнению значений зависимой переменной

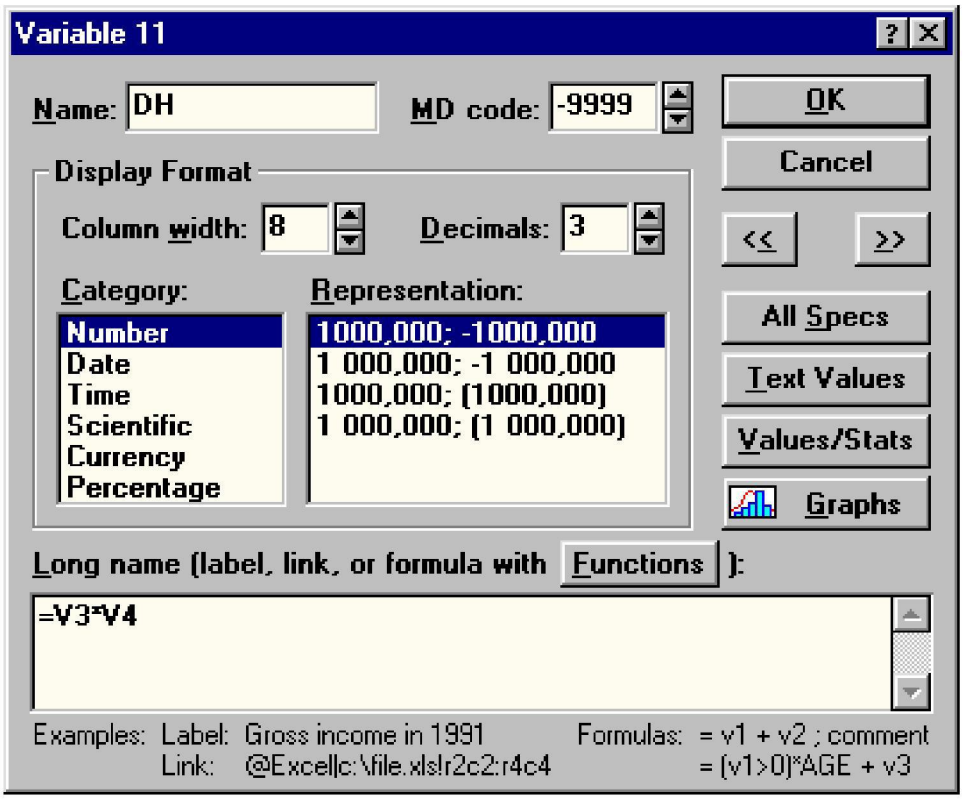


Рис. 10.43. Окно экспликации 11-й переменной

Опцию **Mode** стартового окна регрессионного анализа выставим в положение **Fixed non linear**.

Определим тип преобразования переменных – возведение в квадрат (см. рис. 10.34) и уточним зависимую и независимые переменные модели (рис. 10.44).

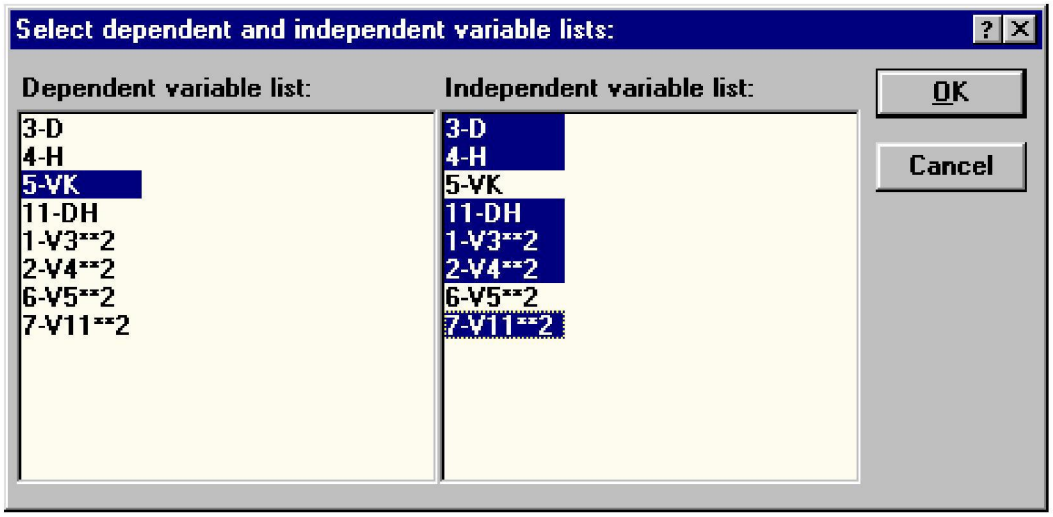


Рис. 10.44. Уточнение зависимой и независимых переменных регрессионного анализа

Для пошаговых методов регрессионного анализа важно установить величину **Tolerance** (толерантность) и величины частного F-критерия для включения в модель (**F to enter**) и исключения из нее (**F to remove**). Установив величину толерантности, мы создаем барьер для включения в модель переменных, толерантность которых меньше установленной. Если величина толерантности переменной мала, то переменная несет малую дополнительную информацию и включение ее в модель не целесообразно. Какая-либо новая независимая переменная, включаемая в модель, может сильно влиять на зависимую переменную, но если она включается в модель после других переменных, она может уже мало влиять на переменную отклика (например, из-за сильной коррелированности с переменными, уже включенными в мо-

дель). По умолчанию в пакете **Statistica** переменная включается в модель, если частный F-критерий больше или равен 1. Численное значение F-критерия для включения никогда не выбирается меньшим, чем численное значение F- критерия для исключения.

Выставим опции окна **Model Definition** так, как показано на рис. 10.45. В результате процедуры пошагового включения переменных в регрессионную модель получено следующее уравнение: VK = 0,0214 + + 0,0009D^2 – 0,0104D + 0,0003$(DH)^2$. Все коэффициенты уравнения значимы на 5-процентном уровне (p-level < 0,05). Это уравнение объясняет 96,4 % (R^2 = 0,964) вариации зависимой переменной (рис. 10.46). Средняя ошибка уравнения составляет 0,02862 $м^3$.

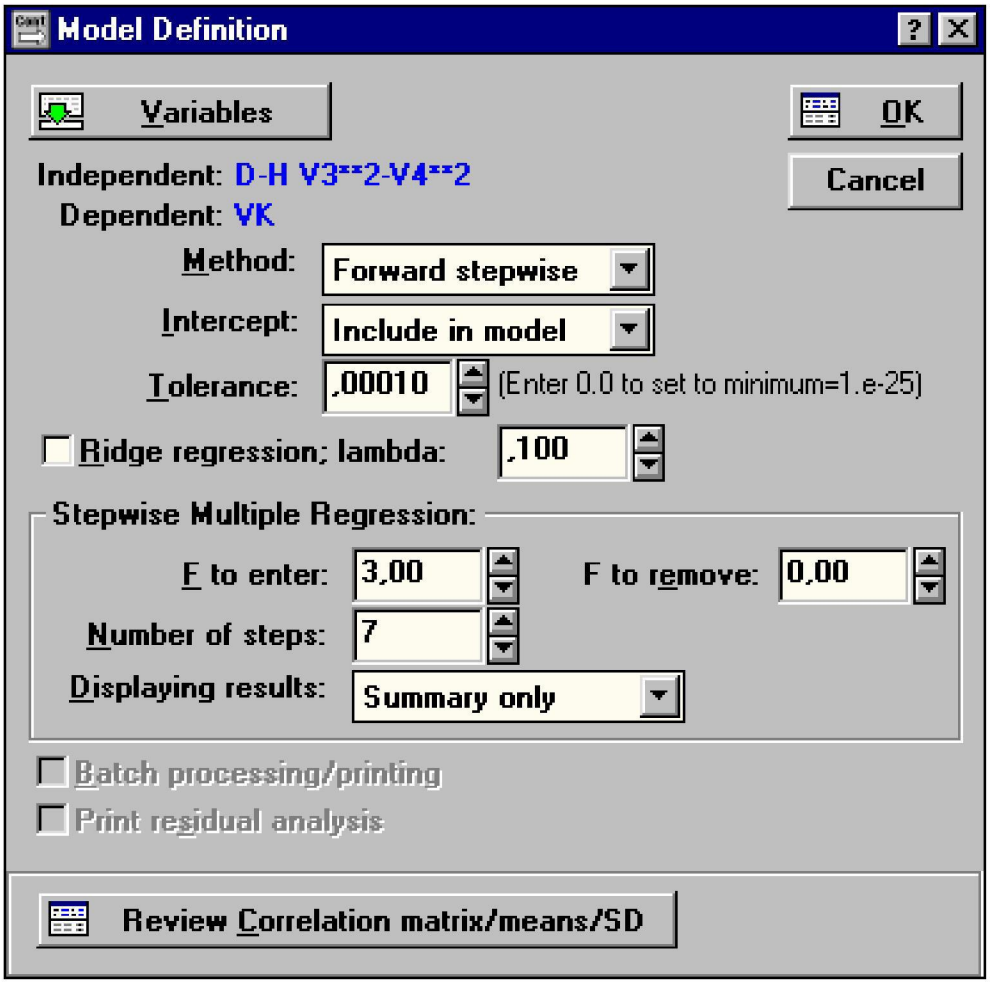
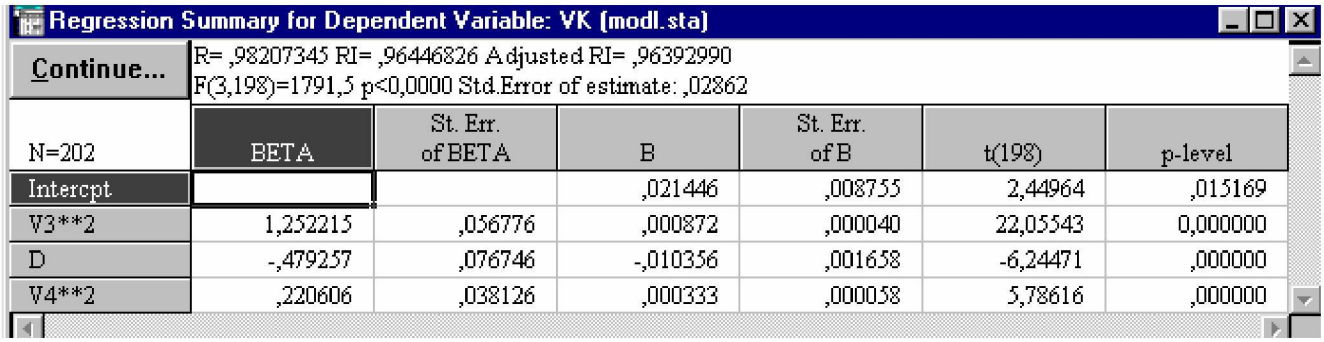


Рис. 10.45. Диалоговое окно Model Definition при использовании метода пошагового включения переменных в модель



Regression Summary for Dependent Variable: VK (modl.sta)

Continue...

R= ,98207345 RI= ,96446826 Adjusted RI= ,96392990
F(3,198)=1791,5 p<0,0000 Std.Error of estimate: ,02862

N=202	BETA	St. Err. of BETA	B	St. Err. of B	t(198)	p-level
Intercpt			,021446	,008755	2,44964	,015169
V3**2	1,252215	,056776	,000872	,000040	22,05543	0,000000
D	-,479257	,076746	-,010356	,001658	-6,24471	,000000
V4**2	,220606	,038126	,000333	,000058	5,78616	,000000

Рис. 10.46. Характеристика уравнения, полученного методом Forward stepwise

При поиске лучшей регрессионной модели необходимо руководствоваться следующими наиболее общими требованиями (Дрейпер, Смит, 1981).

1. Регрессионная модель должна объяснять не менее 80 % вариации зависимой переменной, т. е. $R^2 \geq 0{,}8$.

2. Стандартная ошибка оценки зависимой переменной по уравнению должна составлять не более 5% среднего значения зависимой переменной.

3. Коэффициенты уравнения регрессии и его свободный член должны быть значимы на 5-процентном уровне.

4. Остатки от регрессии должны быть без заметной автокорреляции ($r < 0{,}30$), нормально распределены и без систематической составляющей.

Чем меньше сумма квадратов остатков, чем меньше стандартная ошибка оценки и чем больше R^2 , тем лучше уравнение регрессии.

Одним из недостатков классического регрессионного анализа, в основе которого лежит метод наименьших квадратов, является недостаточная устойчивость к изменениям входной информации. Сейчас довольно широко стали применяться альтернативные регрессионные модели, одной из которых является **гребневая регрессия,** которая отличается устойчивостью для случаев сильной коррелированности зависимых переменных друг с другом. В отличие от метода наи-

меньших квадратов, дающего несмещенные оценки коэффициентов уравнения, в методе гребневой регрессии оценки смещенные, но при этом они имеют меньшую дисперсию. Поэтому такие оценки могут давать более точные и приемлемые для практического использования модели (Забелин, 1983).

Для расчета гребневой регрессии следует установить флажок в опции **Ridge regression** диалогового окна Model Definition.

При практическом использовании метода гребневой регрессии одним из основных вопросов является выбор параметра λ (**lambda)**. Существует несколько численных методов расчета параметра, но чаще используют простой эмпирический подход: выбирают такой параметр λ, при котором коэффициенты стабилизируются и при дальнейшем увеличении параметра изменяются мало. Значение принятого параметра λ является мерой смещения оценок от истинного значения, поэтому стараются не придавать λ слишком больших значений.

Обычно λ выбирают меньше 0,5, а шаг при подборе выбирают небольшим, например 0,02 (Уланова, Забелин, 1990). При λ = 0 уравнение имеет коэффициенты классического метода наименьших квадратов.

10.3. Факторный анализ в системе STATISTICA

Рассмотрим основные этапы проведения кластерного анализа в системе STATISTICA на следующем примере.

Исходными показателями послужили:

X1 – численность населения (тыс.),

X2 – количество человек, приходящихся на одного врача,

X3 –3 Расходы на здравоохранение на душу населения ($) ,

X4 – Уровень детской смертности,

X5 – ВВП, рассчитанный по паритету покупательной способности на душу населения (млн $),

X6 – Смертность на 1000 человек.

В файле (рис. 10.47) содержатся данные по 12 странам (по медицинской характеристике).

Data: Медицинская характеристика (6v by 12c)

	1 Числ_Насел	2 Кол_чел_на_1врача	3 Расх_на_здрав	4 Ур_детск_смерт	5 ВВП_на_душу_насел	6 Смертность
Россия	145491	235	159	16,8	7700	13,9
Азербайджан	8041	256	99	29,3	3000	9,6
Армения	3787	198	152	15,4	3000	9,7
Белоруссия	10187	222	157	12,5	7500	14
Грузия	5262	182	152	17,6	4600	14,6
Казахстан	16172	265	154	42,1	5000	10,6
Киргизия	4921	301	118	37	2700	9,1
Молдавия	4295	251	143	20,5	2500	12,6
Таджикистан	6087	439	100	53,3	1140	8,6
Туркмения	4737	320	125	48,6	4300	9
Узбекистан	24881	299	116	36,7	2400	8
Украина	49568	224	131	15,3	3850	16,4

Рис. 10.47. Исходные данные

Задачей факторного анализа является объединение большого количества показателей, признаков, которыми характеризуется экономический процесс или объект, в меньшее количество искусственно построенных на их основе факторов, чтобы полученная в итоге система факторов (столь же хорошо описывающая выборочные данные, что и исходная) была наиболее удобна с точки зрения содержательной интерпретации.

Алгоритм выполнения

Модуль **Factor Analysis** (факторный анализ) содержит широкий набор методов, снабжающих пользователя исчерпывающими средствами выделения факторов и представления результатов. Для вызова данного модуля можно использовать **STATISTICA Module Switcher**

(Переключатель модулей), который содержит список всех доступных модулей (рис. 10.48), или через меню **Статистика \ Многомерные исследовательские методы \ Анализ фактора (**в зависимости от версии программы STATISTICA**)**.

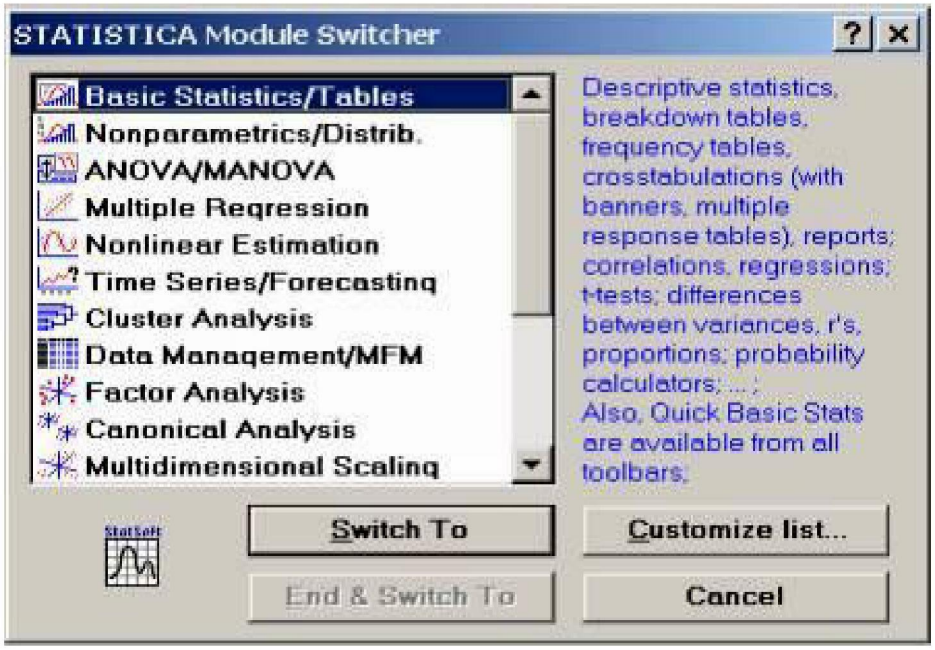


Рис. 10.48. Вид окна **STATISTICA Module Switcher**

На экране появится диалоговое окно (рис. 10.49) **Factor Analysis**:

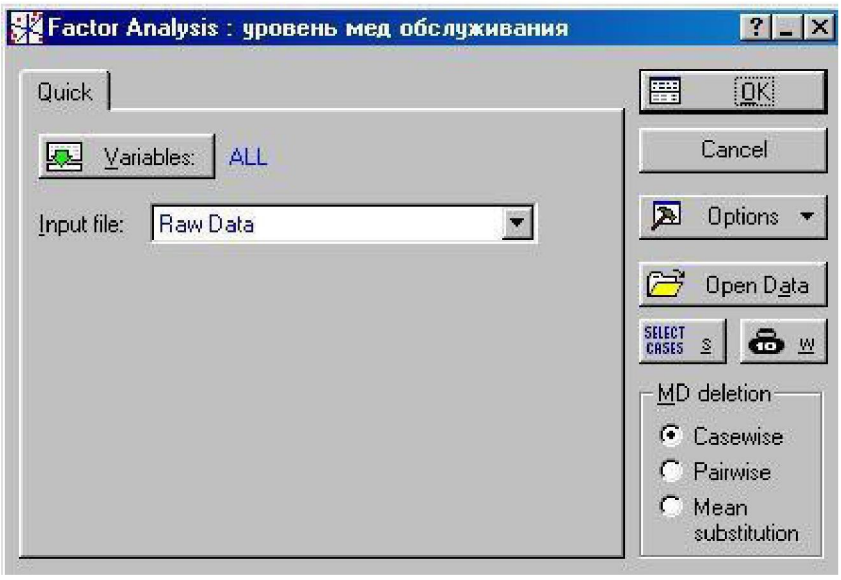


Рис. 10.49. Вид окна **Factor Analysis**

Кнопка **Variables** (Переменные) позволяет отобрать все переменные из файла данных, которые должны быть включены в факторный анализ (рис. 10.50). Если при анализе будут использованы все переменные, то можно воспользоваться кнопкой **Select All** (Выбрать все).

Input File (Входной файл) должен содержать или необработанные данные, или матрицу корреляций и быть предварительно созданным в модуле **Factor Analysis** (Факторный анализ) или другом статистическом модуле.

В модуле возможны следующие типы исходных данных:

- **Correlation Matrix (**Корреляционная матрица);
- **Raw Data** (Исходные данные).

Выберите, например, **Raw Data**. Это обычный файл данных, где по строкам записаны значения переменных.

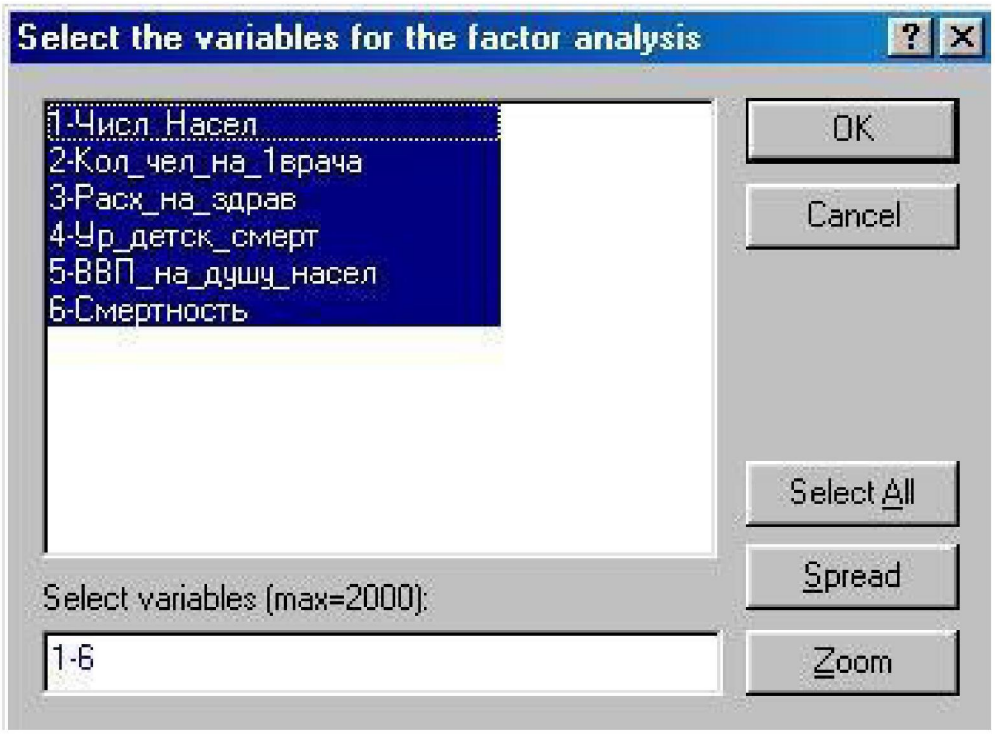


Рис. 10.50. Данные для факторного анализа

MD deletion (замена пропущенных переменных) (рис. 10.49). Способ обработки пропущенных значений:

- **Casewise** (способ исключения пропущенных случаев) – состоит в том, что в электронной таблице, содержащей данные, игнорируются все строки (случаи), в которых имеется хотя бы одно пропущенное значение. Это относится ко всем переменным. В таблице остаются только случаи, в которых нет ни одного пропуска;
- **Pairwise** (парный способ исключения пропущенных значений) – игнорируются пропущенные случаи не для всех переменных, а лишь для выбранной пары. Все случаи, в которых нет пропусков, используются в обработке, например, при поэлементном вычислении корреляционной матрицы, когда последовательно рассматриваются все пары переменных. Очевидно, в способе Pairwise остается больше наблюдений для обработки, чем в способе Casewise. Тонкость, однако, состоит в том, что в способе Pairwise оценки различных коэффициентов корреляции строятся по разному числу наблюдений;
- **Mean Substitution** (подстановка среднего вместо пропущенных значений).

Щелкнув в стартовом окне модуля на кнопку **OK**, вы начнете анализ выбранных переменных.

STATISTICA обработает пропущенные значения тем способом, какой вы ей указали, вычислит корреляционную матрицу и предложит на выбор несколько методов факторного анализа.

Вычисление корреляционной матрицы (если она не задается сразу) – первый этап факторного анализа.

После щелчка по **OK** можно перейти к следующему диалоговому окну.

Define Method of Factor Extraction (определить метод выделения факторов) (рис. 10.51).

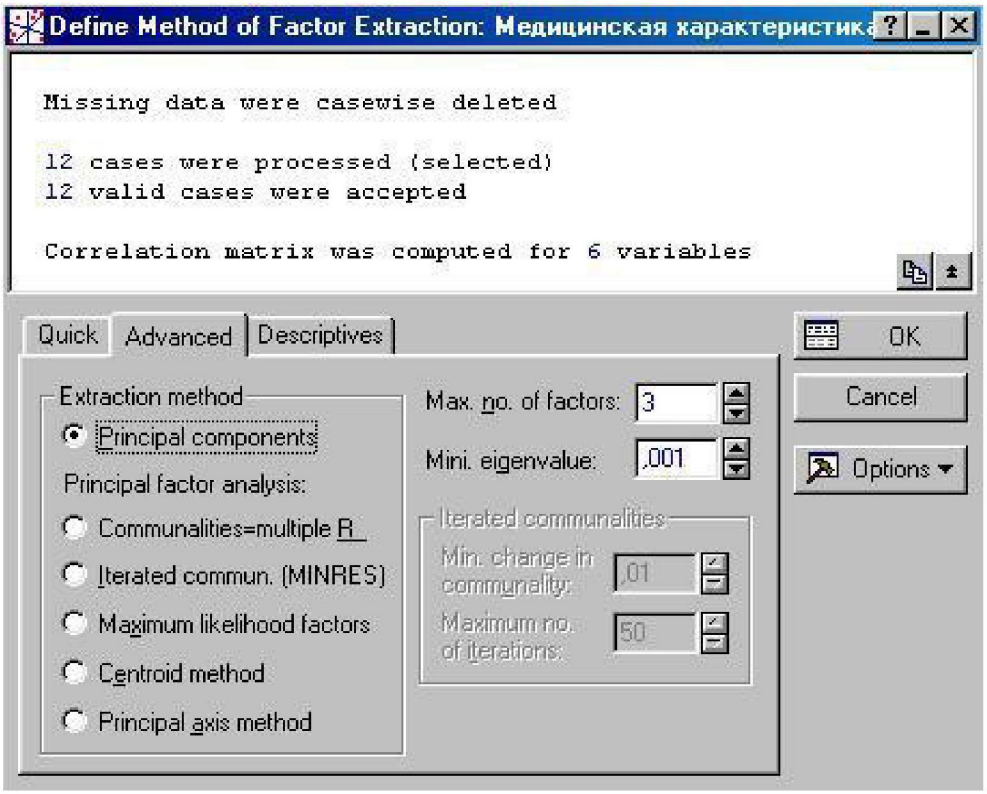


Рис. 10.51. Вид окна **Define Method of Factor Extraction**

Данное окно имеет следующую структуру.

Верхняя часть окна является информационной: здесь сообщается, что пропущенные значения обработаны методом Casewise. Обработано 12 случаев и 12 случаев приняты для дальнейших вычислений. Корреляционная матрица вычислена для 6 переменных.

Группа опций, объединенных под заголовком **Extraction method** (методы выделения факторов) позволяет выбрать метод обработки:

- **Principal components** (метод главных компонент) – позволяет выделить компоненты, работая с первоначальной матрицей корреляций;
- **Communalities=multiple R** (общности как множественный R-квадрат) – на диагонали матрицы корреляций будут находиться оценки квадрата коэффициента множественной корреляции R^2 (соответствующей переменной со всеми другими переменными);

– **Iterated communalities (MINRES)** (метод минимальных остатков) – выполняется в два этапа. Сначала оценки квадрата коэффициента множественной корреляции R^2 используются для определения общностей, как в предыдущем методе. После первоначального выделения факторов метод корректирует их нагрузки с помощью метода наименьших квадратов с целью минимизировать остаточные суммы квадратов;

– **Maximum likelihood factors** (метод максимального правдоподобия) – в этом методе считается заранее известным число факторов (оно устанавливается в поле ввода максимального числа факторов, см. ниже). STATISTICA оценит нагрузки и общности, которые максимизируют вероятность наблюдаемой в таком случае матрицы корреляций. В диалоговом окне результатов анализа доступен χ-квадрат тест для проверки справедливости принятой гипотезы о числе общих факторов;

– **Centroid method** (центроидный метод) – основан на геометрическом подходе;

– **Principal axis method** (метод главных осей) – основан на итеративной процедуре вычисления общностей по текущим собственным значениям и собственным векторам. Итерации продолжаются до тех пор, пока не превышено максимальное число итераций или минимальное изменение в общностях больше, чем это определено в соответствующем поле (см. ниже);

– **Max. no. of factors** (максимальное число факторов). Заданное в этом поле число определяет, сколько факторов может быть выделено при работе рассмотренных выше методов. Это поле работает вместе с полем **Min. eigenvalue** (минимальное собственное значение). Часто при заполнении этого поля руководствуются критерием Кайзера, который рекомендует использовать лишь те факторы, для которых собственные значения не меньше 1.

Остальные поля доступны только при выбранном методе **Centroid method** (центроидный метод) или **Principal axis method** (метод главных осей) и определяют необходимые для успешного выполнения последовательных итераций параметры минимального изменения в общностях и максимального числа итераций.

В окне **Define Method of Factor Extraction** (Определить метод выделения факторов) (рис. 10.52) щелкните на кнопке **Review correlations, means, standart deviations** (Просмотреть корреляции /средние/стандартные отклонения).

Перед вами появилось окно просмотра описательных статистик для анализируемых данных (рис. 10.53), где можно посмотреть средние, стандартные отклонения, корреляции, ковариации, построить различные графики. Здесь можно провести дополнительный анализ текущих данных, проверить соответствие выборочных переменных нормальному закону распределения и существование линейной корреляции между переменными.

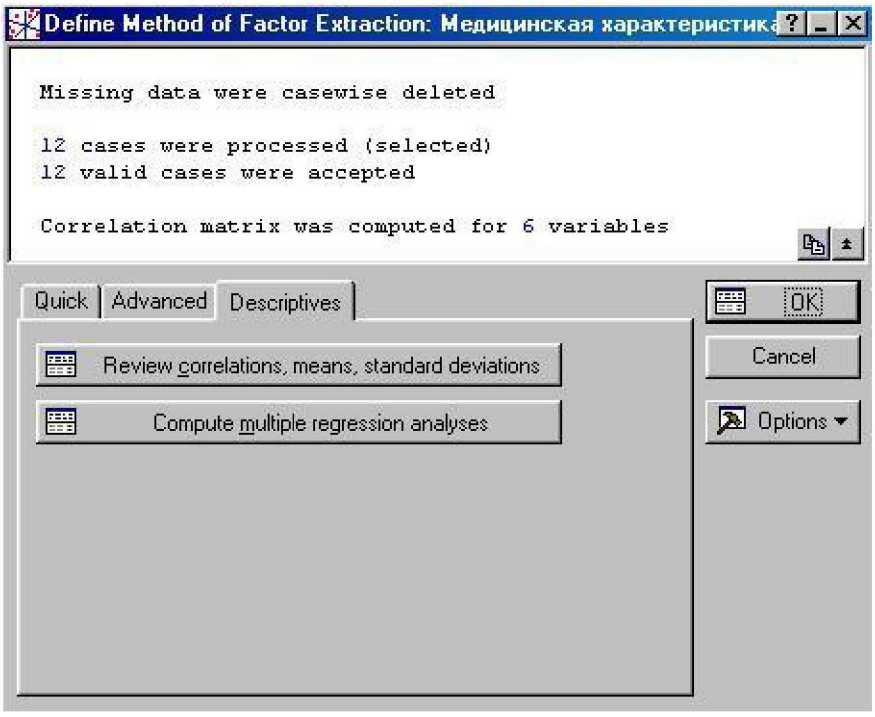


Рис. 10.52. Вид кнопки **Review correlations, means, standart deviations**

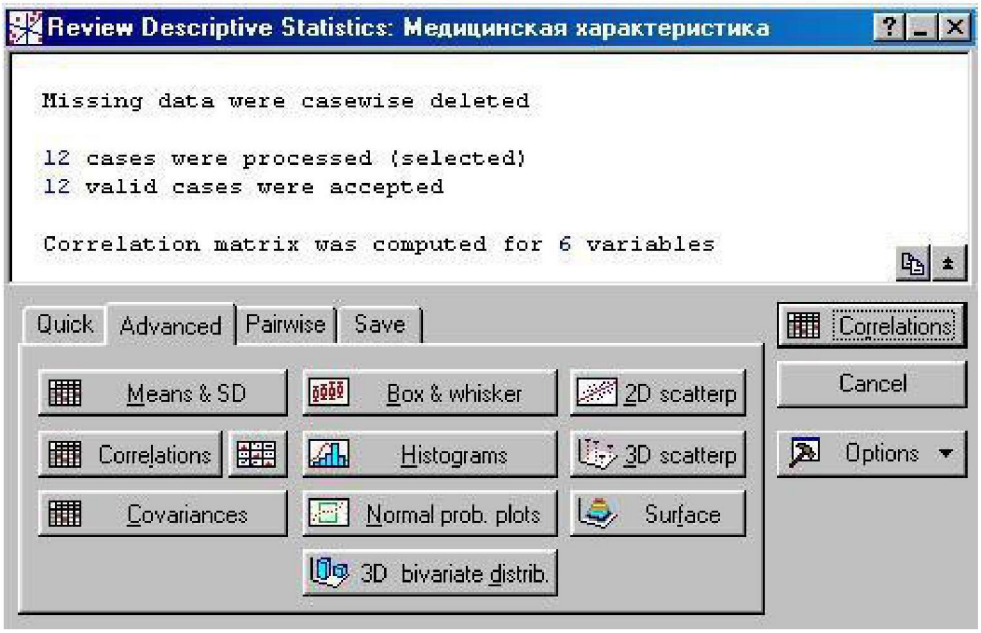


Рис. 10.53. Вид окна просмотра описательных статистик для анализируемых данных

Щелкните на кнопке **Correlations** (**Корреляции**) (рис. 10.54). Вы увидите на экране корреляционную матрицу выбранных ранее переменных.

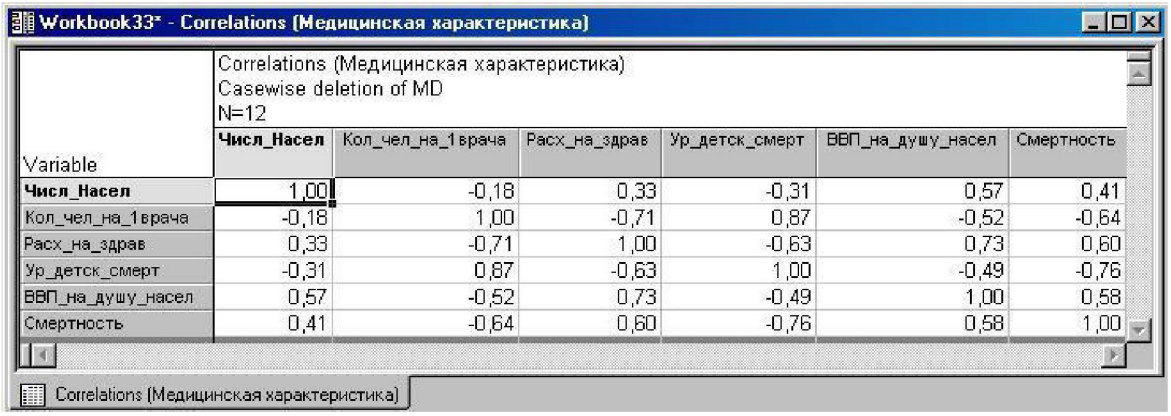


Рис. 10.54. Вид корреляционной матрицы

Выберите опцию **Principal components** (Главные компоненты) и щелкните по кнопке **OK**.

Система быстро произведет вычисления, и на экране появится окно **Factor Analysis Results** (Результаты факторного анализа) (рис.10.55).

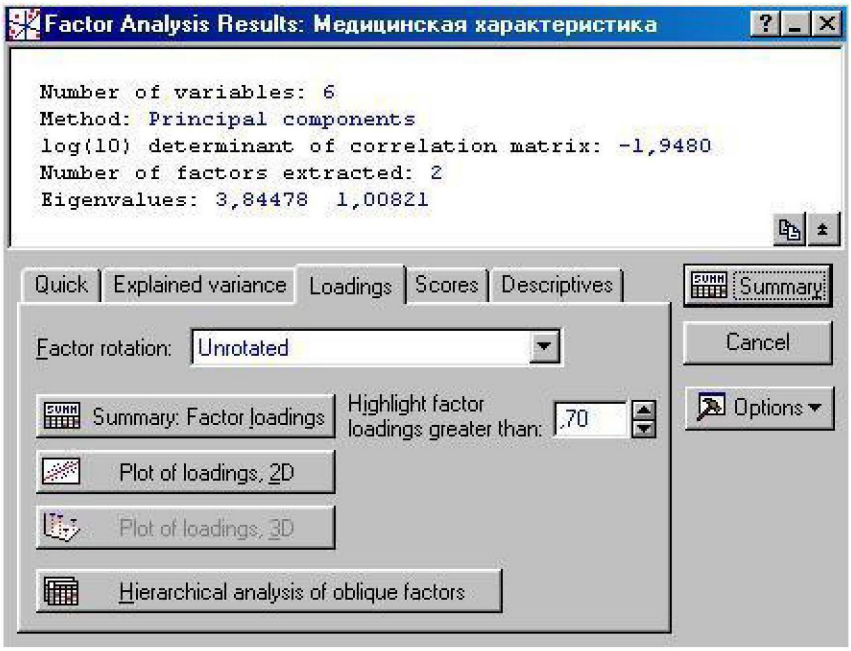


Рис. 10.55. Вид окна **Factor Analysis Results**

В верхней части окна **Результаты факторного анализа** дается информационное сообщение:

Number of variables (число анализируемых переменных) – 6;

Method (метод анализа) – главные компоненты;

log(10) determination of correlation matrix (десятичный логарифм детерминанта корреляционной матрицы) – -1,9480;

Number of Factor extraction (число выделенных факторов) – 3;

Eigenvalues (собственные значения) – 3,84478; 1,00821.

В нижней части окна находятся подразделы, позволяющие всесторонне просмотреть результаты анализа численно и графически:

Plot of loadings, 2D и **Plot of loadings, 3D** (Графики нагрузок) – эти опции построят графики факторных нагрузок в проекции на плоскость любых двух выбранных факторов (рис. 10.56) и в проекции в пространство трех выбранных факторов (для чего необходимо наличие как минимум трех выделенных факторов);

Summary. Factor loadings (Факторные нагрузки). Эта опция вызывает таблицу с текущими факторными нагрузками (рис. 10.57), т. е. вычисленными для данного метода вращения факторов, который указан справа от соответствующей кнопки. В этой таблице факторам со-

ответствуют столбцы, а переменным – строки и для каждого фактора указывается нагрузка каждой исходной переменной, показывающая относительную величину проекции переменной на факторную координатную ось.

Факторные нагрузки могут интерпретироваться как корреляции между соответствующими переменными и факторами: чем выше нагрузка по модулю, тем больше близость фактора к исходной переменной. Таким образом, они представляют наиболее важную информацию для интерпретации полученных факторов. В сгенерированной таблице для облегчения трактовки будут выделены факторные нагрузки по абсолютной величине больше 0,7.

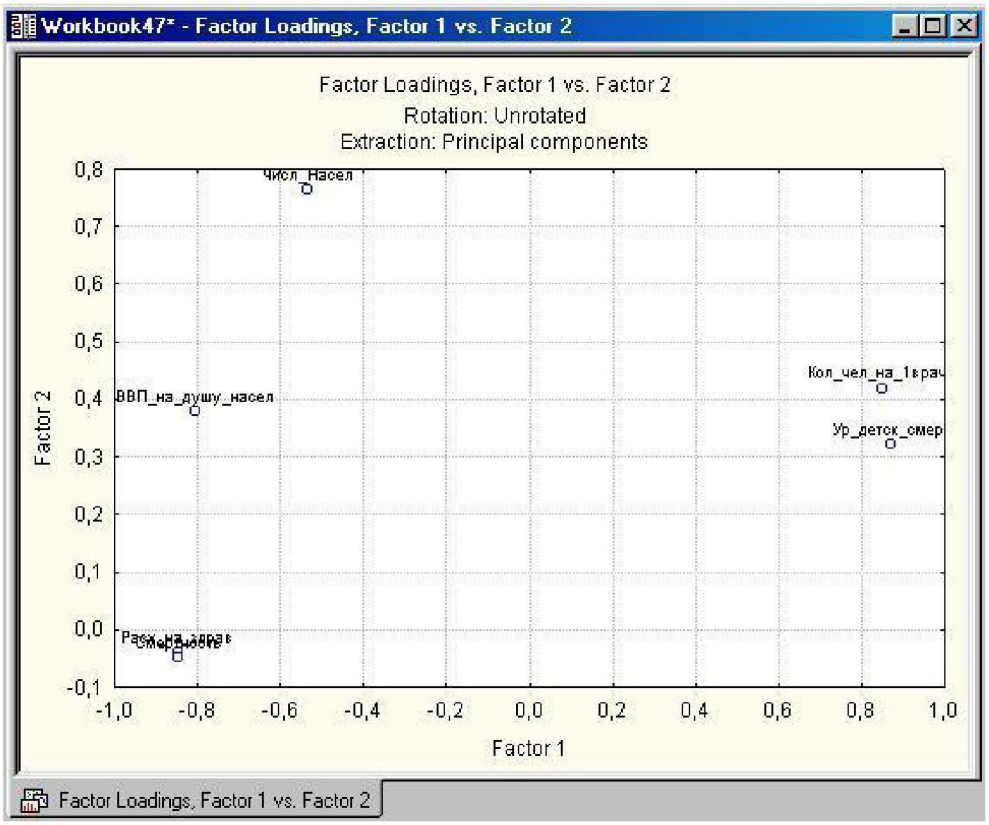


Рис. 10.56. Графики нагрузок

По-видимому, первый фактор более коррелирует с переменными, чем второй. Их трудно проинтерпретировать, возникает вопрос, какой смысл придать второму фактору. В этом случае целесообразно прибегнуть к повороту осей, надеясь получить решение, которое можно интерпретировать в предметной области.

Щелкните по меню **Factor rotation** (Вращение факторов) (рис. 10.58).

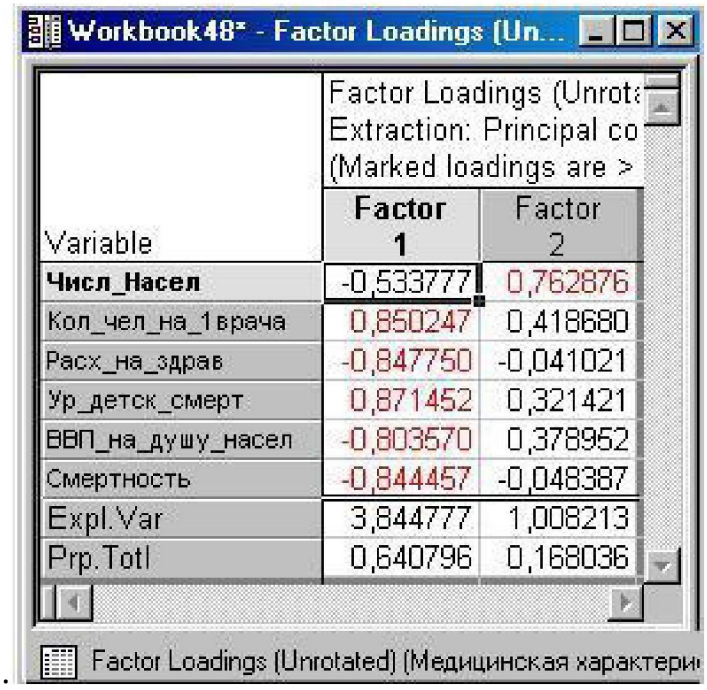


Рис. 10.57. Таблица с текущими факторными нагрузками

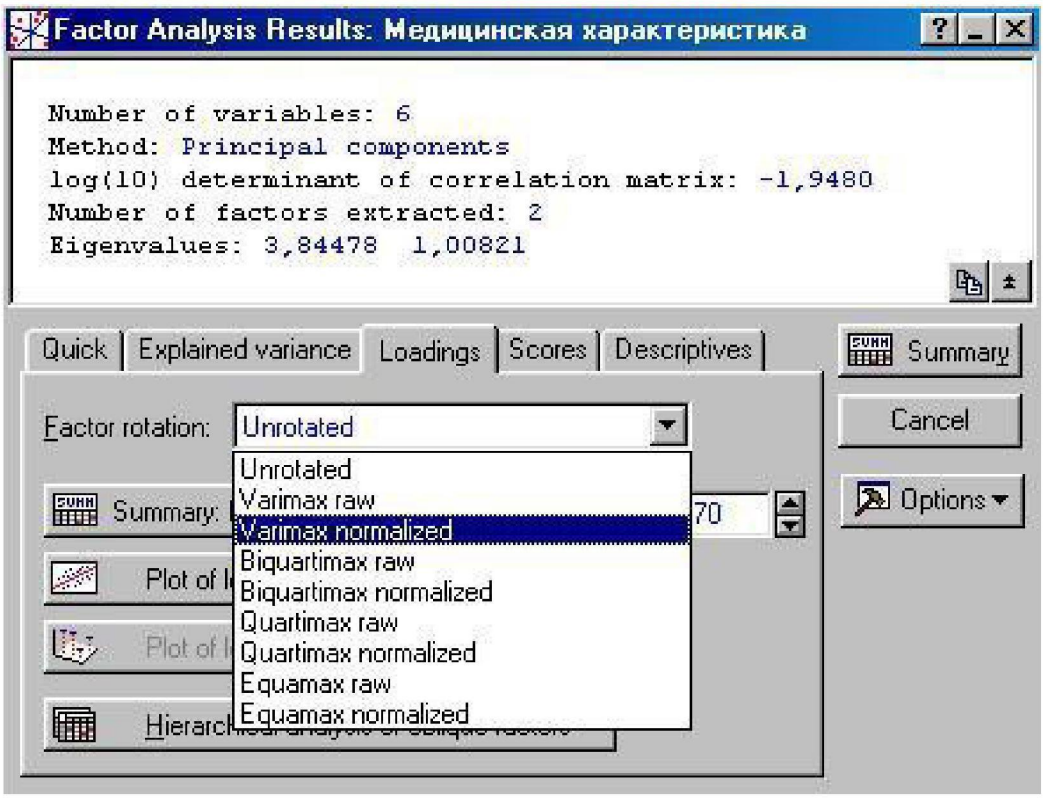


Рис. 10.58. Вращение факторов

Цель вращения – получение простой структуры, при которой большинство наблюдений находится вблизи осей координат. При случайной конфигурации наблюдений невозможно получить простую структуру.

В данном раскрывающемся меню вы можете выбрать различные повороты оси. Окно предлагает несколько возможностей оценить и найти нужный поворот следующими методами:

Varimax – Варимакс;

Biquartimax – Биквартимакс;

Quartimax – Квартимакс;

Equamax – Эквимакс.

Дополнительный термин в названии методов – **normalized** (нормализованные) – указывает на то, что факторные нагрузки в процедуре нормализуются, т. е. делятся на корень квадратный из соответствующей дисперсии. Термин **raw** (исходные) показывает, что вращаемые нагрузки не нормализованы.

Инициируйте кнопку **Varimax normalized** (Варимакс нормализованный).

Система произведет вращение факторов методом нормализованного Варимакса, и окно **Factor Analysis Results** (Результаты факторного анализа) снова появится на мониторе. Вновь инициируйте в этом окне кнопку **Plot of Loadings 2D** (Двумерный график нагрузок). Вы опять увидите график нагрузок (рис. 10.59).

Конечно, этот график немного отличается от предыдущего. Посмотрим еще нагрузки численно, инициировав кнопку **Факторные нагрузки** (Factor loadings). Щелкните на кнопке **Summary. Factor loadings** и откроется окно (рис. 10.60).

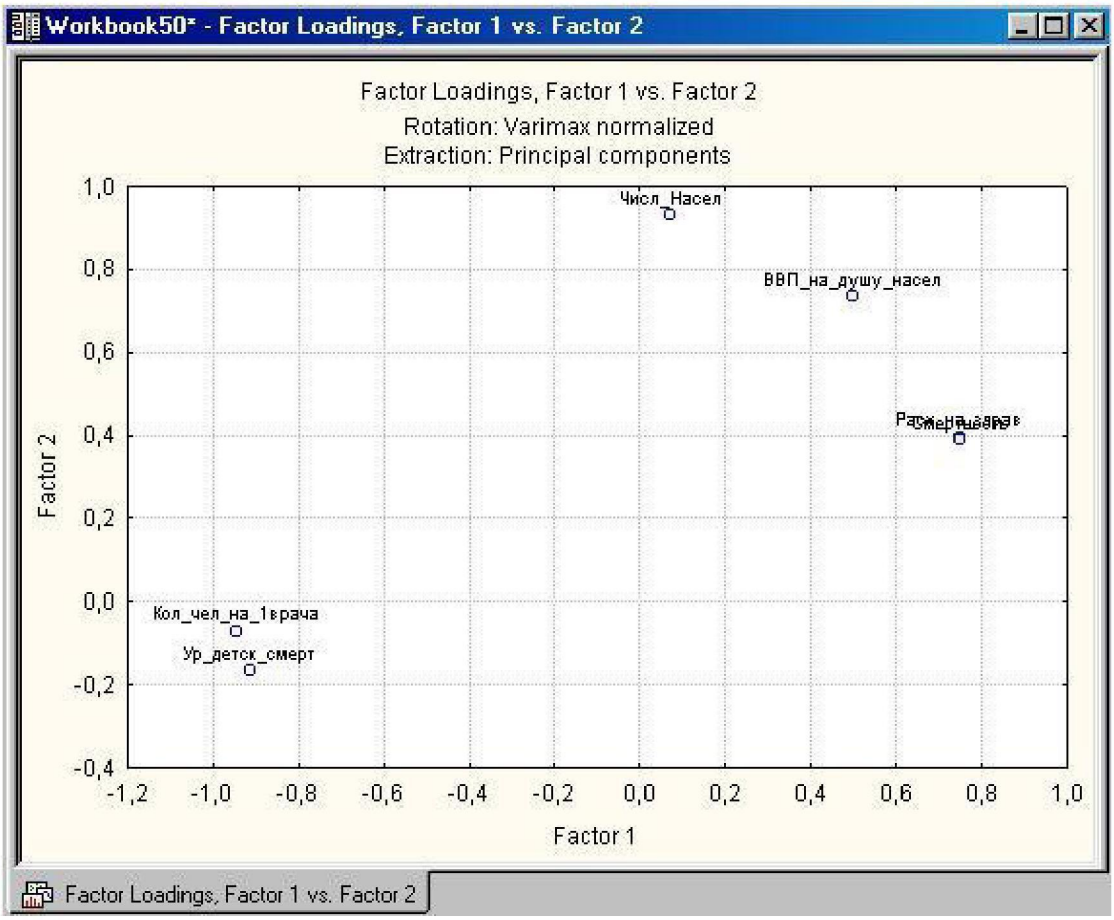


Рис. 10.59. Двумерный график нагрузок

Рис. 10.60. Окно таблиц

Теперь найденное решение уже можно интерпретировать. Факторы чаще интерпретируют по нагрузкам. Первый фактор теснее всего связан с X2 , X3, X4, X6. Второй фактор – X1 и X5. Таким образом, произвели классификацию переменных на две группы. Возникает вопрос: сколькими же факторами следует ограничиваться на практике? Для этого в программном пакете STATISTICA существует критерий **Scree plot** (Критерий каменистой осыпи). В окне **Factor Analysis Results** нажмите кнопку **Scree plot**, получите следующий график собственных значений (рис.10.61).

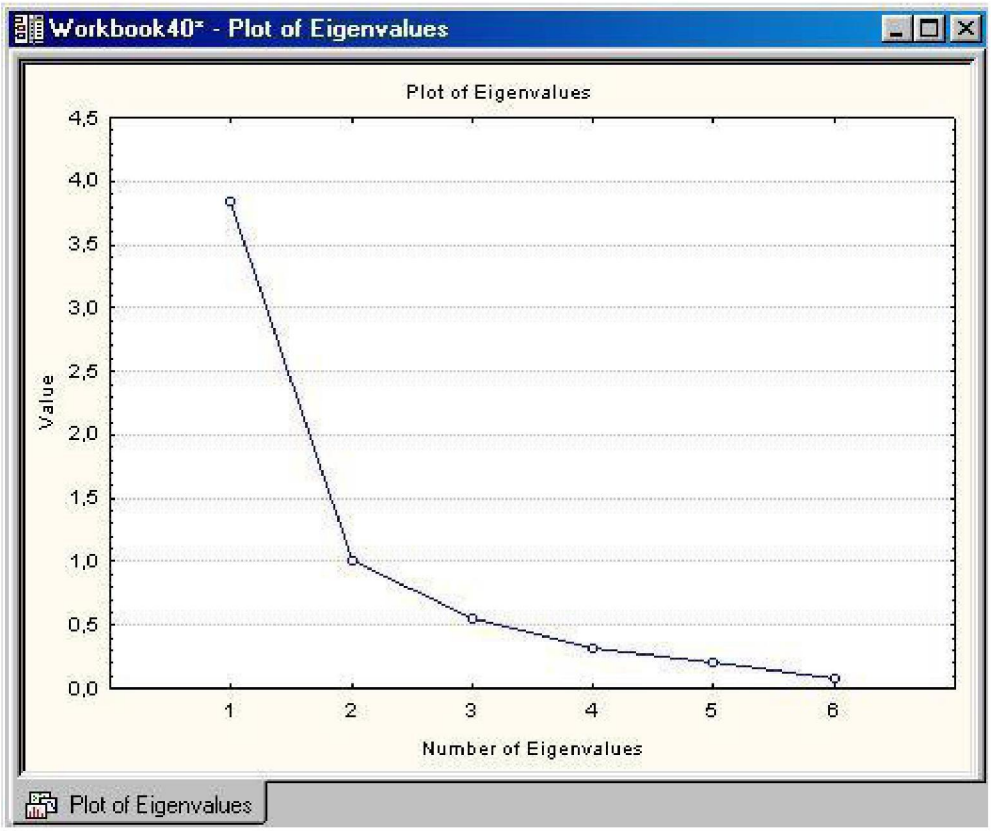


Рис. 10.61. График собственных значений

В точках с координатами 1, 2 осыпание замедляется наиболее существенно, следовательно, теоретически можно ограничиваться двумя факторами.