МИНИСТЕРСТВО СЕЛЬСКОГО ХОЗЯЙСТВА И ПРОДОВОЛЬСТВИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ АГРАРНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

КАФЕДРА МОДЕЛИРОВАНИЯ И ПРОГНОЗИРОВАНИЯ ЭКОНОМИКИ АПК

СТАТИСТИЧЕСКАЯ ОБРАБОТКА ИНФОРМАЦИИ С ПОМОЩЬЮ ПАКЕТА «STATISTICA»

Учебно-методическое пособие

Минск
2008

УДК 004.4:311(0175.8)
ББК 32.81я7
С 78

Рекомендовано научно-методическим советом факультета предпринимательства и управления БГАТУ

Протокол № 4 от 27 марта 2008 г.

Составитель – д.т.н., проф. *В.А. Грабауров*

Рецензенты: к.т.н., доцент, зав. каф. экономической информатики БГАТУ *В.С. Ионин*;
к.э.н., доцент каф. экономической информатики БГЭУ *В.Н. Гулин*

Статистическая обработка информации с помощью пакета «Statistica» : С 78 учеб-метод. пособие / сост. В.А. Грабауров. – Минск : БГАТУ, 2008. – 96 с.
ISBN 978-985-519-014-2

В учебно-методическом пособии излагается методика обработки экспериментальных данных с помощью компьютерной программы «Statistica for Windows».

Предназначено для студентов экономических специальностей, может быть использовано студентами и аспирантами других специальностей для обработки экспериментальных данных.

УДК 004.4:311(0175.8)
ББК 32.81я7

ISBN 978-985-519-014-2

Содержание

Введение 5

1 Основные понятия и особенности ППП Statistica for Windows 6

1.1 Многомерное представление экономического объекта 6

1.2 Структура окон STATISTICA for WINDOWS 8

2 Выполнение регрессионного анализа 13

2.1 Основы метода наименьших квадратов 13

2.2 Нахождение коэффициентов линейного уравнения путем прямого расчета 16

2.3 Нахождение коэффициентов линейного уравнения с использованием ППП STATISTICA for WINDOWS. 17

3 Построение многомерных регрессионных моделей 23

3.1 Построение линейной двумерной модели 25

3.2 Построение квадратичной двумерной модели 31

3.3 Построение квадратичной трехмерной модели 33

3.4 Определение оптимальной структуры модели 35

4 Оценка чувствительности Вашей фирмы к … 36

4.1 Исследование анализа чувствительности 36

4.2 Указания к выполнению 36

5 Куда движется Ваша фирма? 39

5.1 Выделение тренда из экспериментальных данных 39

5.2 Указания к выполнению 40

6 От каких факторов зависит успех Вашей фирмы? 43

6.1 Основы многомерного факторного анализа 43

6.2 Порядок выполнения 45

6.3 Указания к выполнению 47

7 Каким фирмам можно доверять? 53

7.1 Основы кластерного анализа 53

7.2 Порядок выполнения 58

7.3 Указания к выполнению 59
8 К какой группе относится объект? (Постановка диагноза) 70
8.1 Основы дискриминантного анализа 70
8.2 Порядок выполнения 71
8.3 Указания к выполнению 72
9 Эффективны ли нововведения? 85
9.1 Основы однофакторного дисперсионного анализа 85
9.2 Порядок выполнения 87
9.3 Указания к выполнению 90
Список литературы 95

ВВЕДЕНИЕ

Цель работы – выработать практические навыки по проведению статистического анализа экономической информации с использованием персональных компьютеров. В ходе выполнения лабораторных работ используется стандартный пакет прикладных программ ***STATISTICA for WINDOWS***. Пакет ***STATISTICA for WINDOWS*** является очень мощным статистическим пакетом, предназначенным как для начинающего пользователя, так и для профессионального статистика.

В нашей работе мы познакомимся только с наиболее полезными для экономистов приемами статистической обработки, при этом ограничимся лишь первичным анализом результатов. Описываемые методы могут быть использованы для обработки практически любого вида информации, а не только экономической. Все изучаемые методы статистической обработки данных имеют четкую практическую направленность: сначала формулируются задачи исследования, а затем рассматриваются методы и соответствующие разделы пакета ***STATISTICA for WINDOWS***, которые позволяют решить эти задачи.

Предполагается, что исследователь уже имеет первичные знания о математической статистике. Тем не менее, в некоторых разделах рассказывается об основах используемых статистических методов. При этом главное внимание уделяется не доказательству корректности тех или иных приемов, а их содержательной базе. Это сделано для того, чтобы пользователь понимал суть выполняемых процедур.

1 Основные понятия и особенности ППП Statistica for Windows

1.1 Многомерное представление экономического объекта

Каждый экономический объект, будь то предприятие, регион, страна, группа стран типа СНГ или Европейский союз, могут быть охарактеризованы набором показателей, который меняется во времени. Например, на рисунке 1.1 представлены социальные и демографические показатели Республики Беларусь по данным Международного Валютного Фонда.

	2001	2002	2003	2004	2005
Social and demographic indicators					
Area (km)	207,600	207,600	207,600	207,600	207,600
Population (in thousands)	9,951	9,899	9,849	9,800	9,751
Urban	7,031	7,037	7,045	7,056	7,059
Rural	2,920	2,862	2,804	2,744	2,692
Population density (inhabitants per sq. km.)	48	48	47	47	47
Life expectancy at birth (in years)	69.4	68.9	69.0	69.5	68.8
Infant mortality rate (per thousand)	9.1	7.8	7.7	6.9	6.4
Annual population growth rate (in percent)	-0.4	-0.5	-0.5	-0.5	-0.5
GDP (in millions of U.S. dollars)	12,094	14,489	17,622	23,102	29,549
GDP per capita (in U.S. dollars)	1,215	1,464	1,789	2,357	3,030

Рисунок 1.1 Показатели Республики Беларусь по данным Международного Валютного Фонда

Количество таких показателей может быть различным. На рисунке 1.2 представлены макроэкономические показатели США. В зависимости от особенностей исследуемого объекта и цели исследования число показателей может варьироваться в широких пределах. Часть показателей может рассматриваться как входные факторы, действующие на объект управления, другие – как выходные параметры, по которым может оцениваться состояние объекта управления.

Рисунок 1.2 Макроэкономические показатели США

Экономический объект управления (фирма, предприятие) представляется в виде (рисунок 1.3), который характеризуется входными и выходными параметрами (факторами). Иногда деятельность экономического объекта можно оценивать по одному выходному обобщенному параметру, например, прибыли. В качестве входного параметра может использоваться время.

где $\mathbf{X_1}...\mathbf{X_n}$ - входные параметры (факторы), **t** - время,

$\mathbf{Y_1}...\mathbf{Y_m}$ - выходные параметры

Рисунок 1.3 Общий вид объекта управления

1.2 Структура окон STATISTICA for WINDOWS

После запуска программы появится главное окно (рисунок 1.4). Некоторые элементы этого окна такие же, как и у всех других программ ***WINDOWS***, другие присущи только программе ***STATISTICA for WINDOWS***. Остановимся на специфических элементах окна.

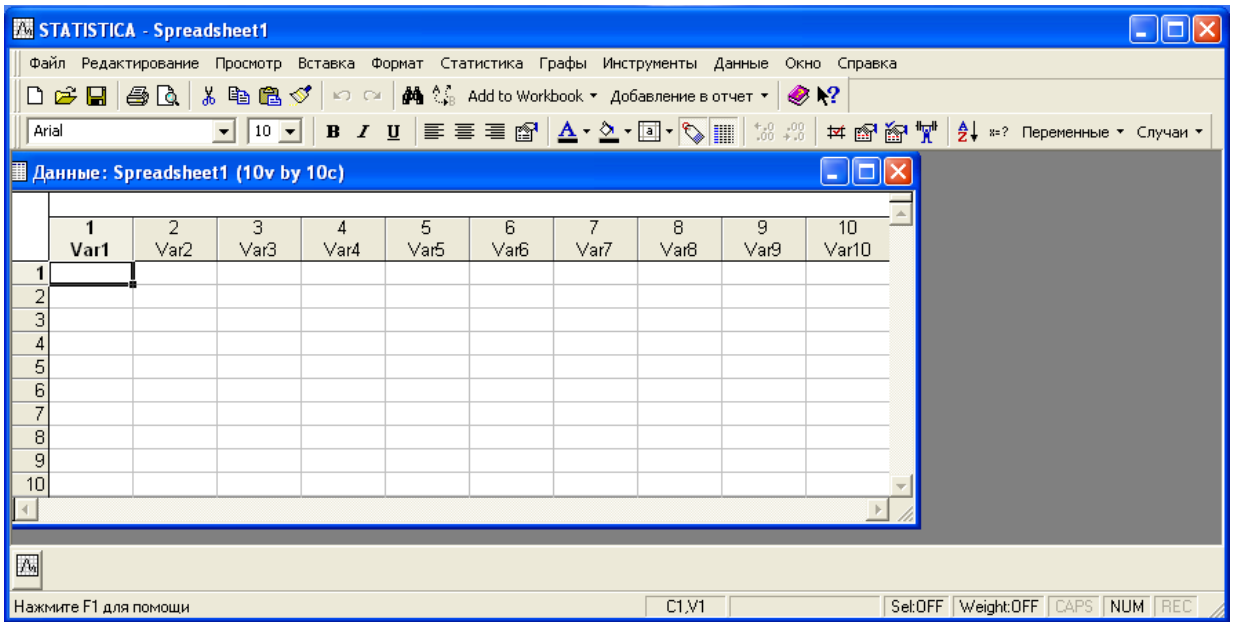


Рисунок 1.4 Главное окно ППП ***STATISTICA for WINDOWS***

Прежде всего, бросаются в глаза особенности окна ***STATISTICA for WINDOWS***: *Меню* (аналогично окнам ***WORD, EXCEL***) и *Таблица данных*.

Меню

Меню содержит традиционный для ***WINDOWS*** набор (*Файл, Редактирование, Просмотр, Вставка, Формат, Сервис, Окно*), а также *Статистика, Графы*, *Данные*, поэтому остановимся только на специфичных и наиболее полезных разделах.

Вставка: добавление и копирование переменных и случаев.

Статистика

При открытии падающего окна *Статистика* появляется список основных статистических процедур, внутри которого встроено множество видов анализа (рисунок 1.5).

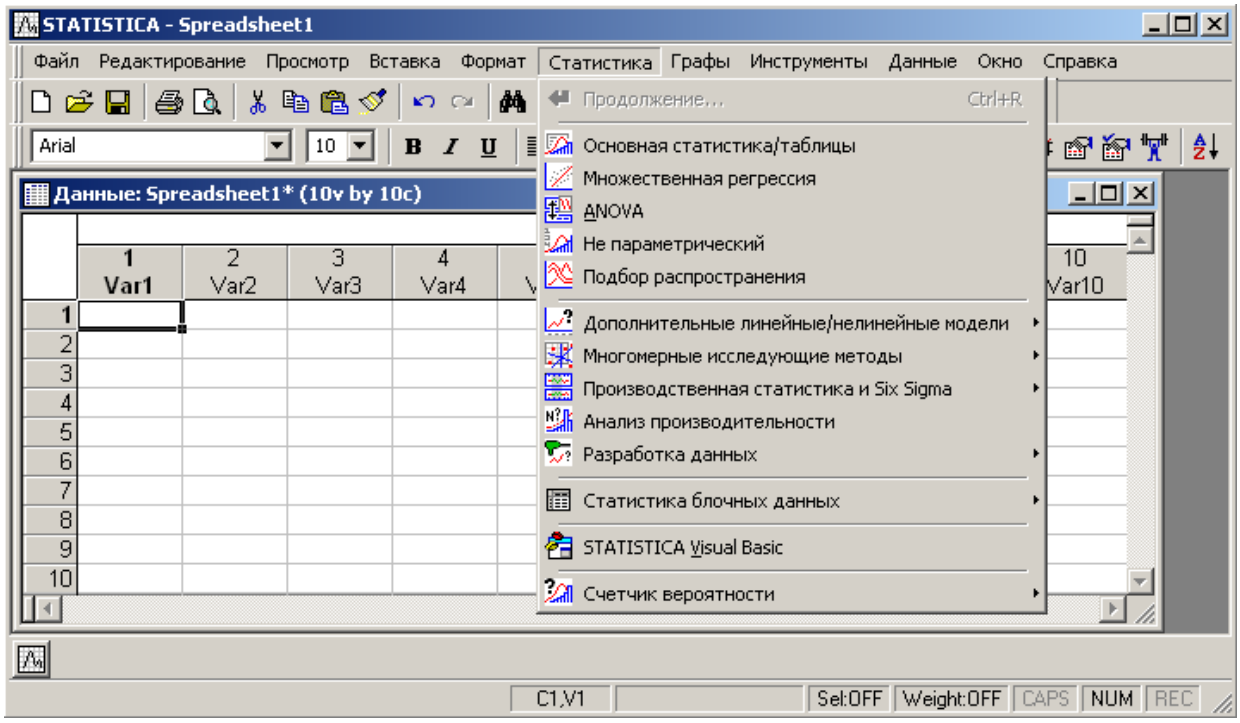


Рисунок 1.5 Окно *Статистика*

Графы

Пакет ***STATISTICA for WINDOWS*** позволяет строить множество видов графиков, как двумерных, так и трехмерных. Эти графики можно преобразовывать и даже вращать вдоль трех осей. Основной список графиков открывается с помощью падающего окна *Графы* (рисунок 1.6).

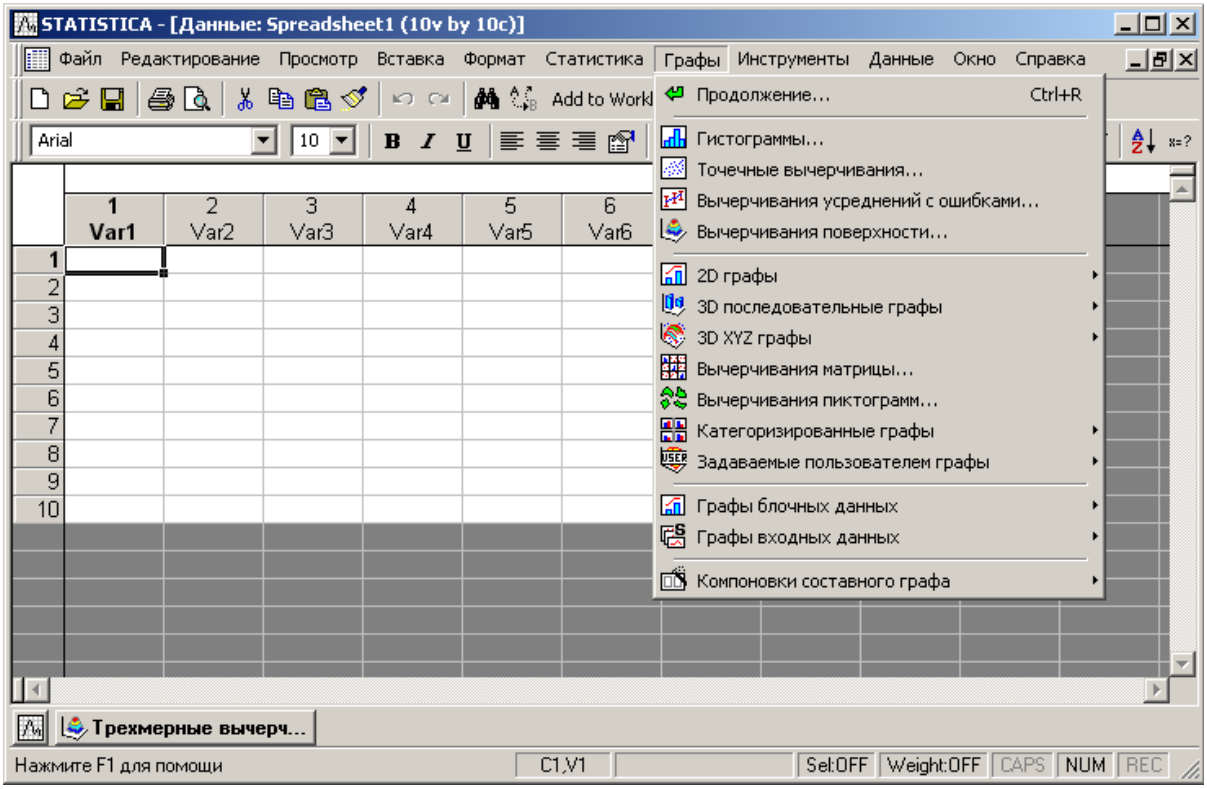


Рисунок 1.6. Окно *Графы*

Таблица данных

Таблица данных состоит из переменных (колонок) и случаев (строк). По умолчанию названия переменных: Var1, Var2, Var3, Var4 и т.д. Эти названия можно задавать произвольно. Для изменения имени переменной нужно двойным щелчком на соответствующей переменной Var вызвать окно спецификации переменной (рисунок 1.7). В этом окне в ячейке *Name* можно задать нужное имя переменной (например, Years). В появившейся по умолчанию таблице данных 10 строк и 10 колонок. При желании можно задать количество строк и колонок произвольно командами *Файл/Новый/Электронные таблицы*.

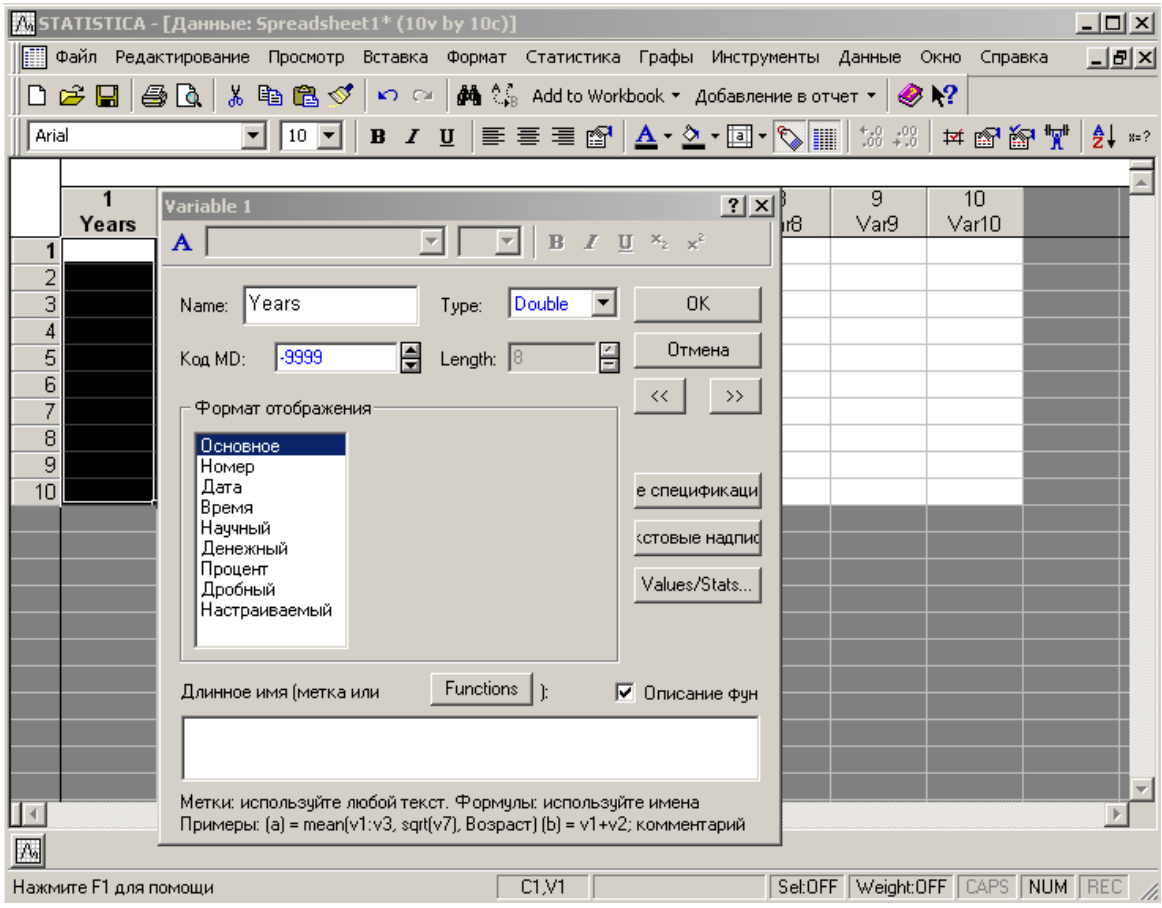


Рисунок 1.7. Окно спецификации переменной

Создание файла набора данных

Файлы набора данных можно создавать самим, импортировать из других пакетов, комбинировать из нескольких и т.д. Здесь мы рассмотрим только действия по созданию нового файла данных.

Если нужная таблица отличается от появляющейся по умолчанию (10*10), можно создать свою таблицу. Для этого выполним следующую последовательность действий: *Файл/Новый/Электронные таблицы*, в результате чего появится окно с параметрами файла - создаваемой таблицы данных (рисунок 1.8). В этом окне количество обозначает число переменных, а число регистров – число случаев.

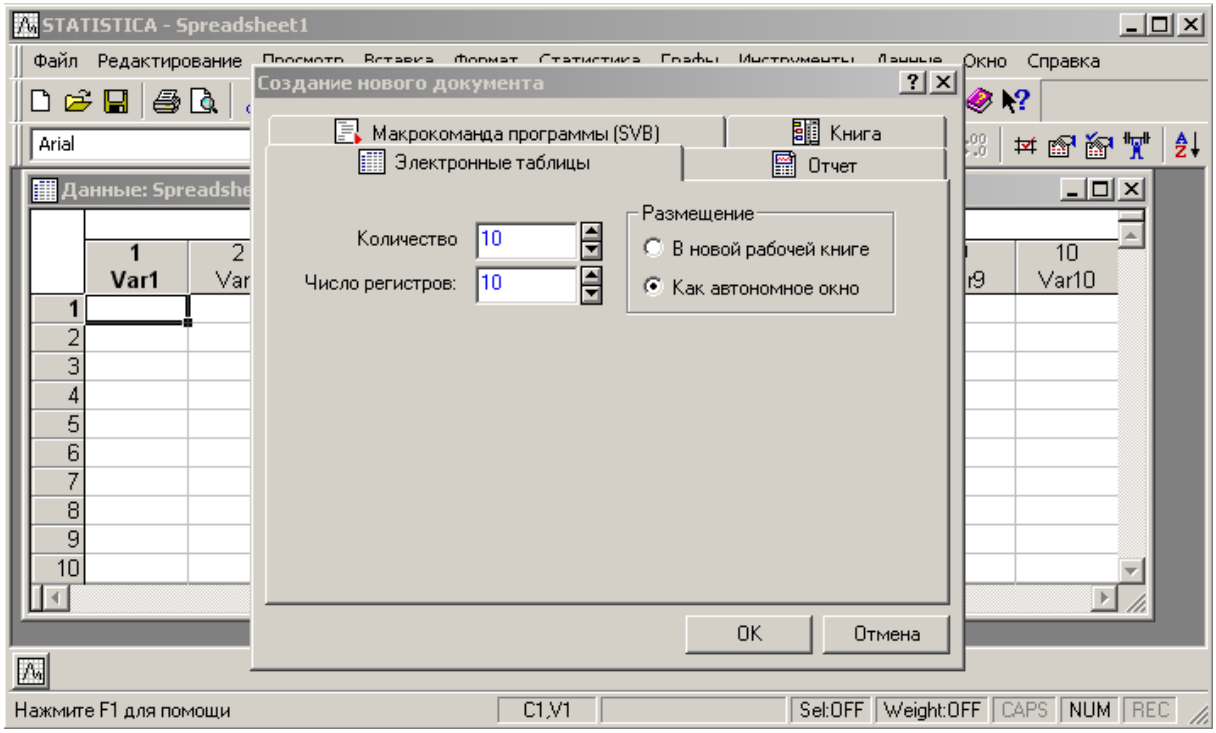


Рисунок 1.8. Окно *Создание нового документа*

2 Выполнение регрессионного анализа

2.1 Основы метода наименьших квадратов

Пакет ***STATISTICA for WINDOWS*** позволяет выполнять множество видов статистических расчетов по встроенным процедурам, часть из которых видна в падающем окне *Статистика* (рисунок 1.5), и еще многие из них появляются при открытии других падающих окон внутри окна *Статистика*. Такие возможности легко выполнять сложные статистические расчеты, с одной стороны, вдохновляют, но, с другой стороны, существует опасность профанации, когда пользователь что-то там считает, не очень понимая, что он делает. Чтобы избежать этой опасности, нужно разобраться в сути производимых статистических расчетов. Помимо этого желательно обладать уверенностью, что имеющиеся в пакете ***STATISTICA for WINDOWS*** процедуры расчета соответствуют заложенным в них теоретическим основам.

Все виды статистических расчетов мы проверить не сможем, поэтому выберем для пробы расчет простейшего уравнения линейной регрессии. Проведем расчет этого уравнения дважды: вручную (или с помощью пакета **Excel**) и с помощью пакета ***STATISTICA for WINDOWS.*** Тем самым мы, во-первых, выясним достоверность заложенных в пакете ***STATISTICA for WINDOWS*** процедур расчета, а, во-вторых, убедимся, насколько облегчает работу пакет ***STATISTICA for WINDOWS*** даже на простейших расчетах, не говоря уже о более сложных.

Предположим, имеется двумерный набор данных (рисунок 2.1). Нашей задачей является проведение линии регрессии (тренда), которая соответствует тенденции. Вопрос заключается в том, как провести линию регрессии, т.е. будет ли она линейной или нелинейной и где она должна проходить.

x	y
x_1	y_1
x_2	y_2
·	·
·	·
·	·
x_n	y_n

Рисунок 2.1 Набор экспериментальных значений

На рисунке 2.2 показан в двумерных координатах этот набор точек и линия регрессии. Каждой точке соответствует ее реальное значение y_i и расчетное значение $\bar{y}_i$, находящееся на линии регрессии. Разница между ними называется отклонением $e_i = y_i - y_i'$. Линию регрессии нужно провести таким образом, чтобы сумма квадратов всех отклонений была минимальной.

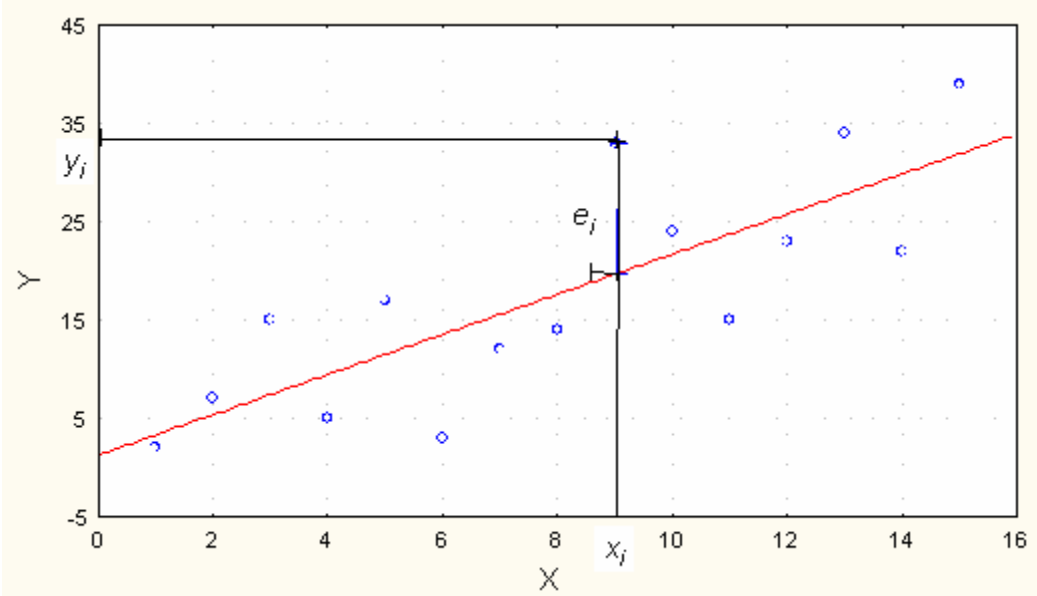


Рисунок 2.2 Набор экспериментальных значений и линия регрессии

Уравнение простейшей линейной функции:

$$\bar{y} = a + bx\,, \qquad (2.1)$$

где x – входная переменная, $\bar{y}_i$ - выходная переменная (расчетное значение).

Нахождение параметров уравнения (a, b) осуществим методом регрессионного анализа, т.е. анализа, связанного с изучением зависимости одной случайной величины от других случайных величин, связанных с нею. Критерием проведения кривой (в данном случае прямой) уравнения регрессии среди массовых экспериментальных точек используем введение К.Гауссом требование минимизации суммы квадратов отклонений S экспериментальных значений y_i от расчетной кривой, т.е. (среднеквадратичной ошибки):

$$S = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 \to \min \qquad (2.2)$$

Этот принцип используется для отыскания не только уравнения прямой линии, но и для построения любой линии регрессии.

Покажем, как надо использовать метод наименьших квадратов на примере нахождения параметров линейного уравнения $\bar{y} = a + bx.$

Начнем с преобразования уравнения (2.2):

$$S = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 = \sum\ [y^2 - 2y(a + bx) + (a + bx)^2];$$

Опуская для краткости индекс i, с учетом того, что $\sum_{i=1}^{n} = n$, получим:

$$S = \sum y^2 - 2a\sum y - 2b\sum xy + a^2 \sum_{i=1}^{n} + 2ab\sum x + b^2 \sum x^2; \qquad (2.3)$$

Нас интересует нахождение таких значений параметров a и b, которые обращают в минимум (2.3). Как известно, для нахождения экстремума необходимо взять частные производные $\partial S/\partial a$ и $\partial S/\partial b$ и приравнять их нулю.

$$\begin{cases} \partial S/\partial a = -2\sum y + 2na + 2b\sum x = 2(na + b\sum x - \sum y) = 0; & (2.4) \\ \partial S/\partial b = -2\sum xy + 2a\sum x + 2b\sum x^2 = 2(a\sum x + b\sum x^2 - \sum xy) = 0. & (2.5) \end{cases}$$

Из (2.4) и (2.5) получаем систему двух уравнений (2.6), из которой определяются параметры a и b.

$$\begin{cases} an + b\sum x = \sum y; \\ a\sum x + b\sum x^2 = \sum xy, \end{cases} \qquad (2.6)$$

Итак, для нахождения искомых параметров a и b необходимо вычислить значения $\sum x$, $\sum y$, $\sum x^2$, и $\sum xy$.

Адекватность модели оценивается с помощью коэффициента множественной корреляции

$$R = \sqrt{1 - \frac{S_e^2}{S_y^2}} \quad , \qquad (2.7)$$

где: S_e^2 - дисперсия отклонения, S_y^2 - дисперсия случайной величины.

Значения коэффициента множественной корреляции могут варьироваться в пределах от 0 до 1. При $R \approx 0$ дисперсия отклонения сравнима с дисперсией случайной величины, а при $R \approx 1$ все экспериментальные точки лежат вблизи линии регрессии.

2.2 Нахождение коэффициентов линейного уравнения путем прямого расчета

Проведем вручную, с помощью калькулятора или пакета ***Excel*** расчет параметров уравнения регрессии вида $y = a + bx$ по системе уравнений (2.6), для чего составим таблицу 2.1 экспериментальных данных размерностью, например, 2*20 ($n = 20$) и рассчитаем по ней значения xy и x^2. (В качестве независимой переменной x используем порядковый номер данных, которые должны быть расположены по мере увеличения). Так как для решения системы уравнений (2.6) нужно иметь помимо значений $\sum x$ и $\sum y$ также значения $\sum x^2$ и $\sum xy$, добавим в таблицу 2.1 колонки значений xy и x^2 .

Таблица 2.1 Экспериментальные и расчетные данные

x	y	xy	x^2
$x_1 = 1$	y_1	$x_1 y_1$	$x_1 x_1$
$x_2 = 2$	y_2	$x_2 y_2$	$x_2 x_2$
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
$x_n = n$	y_n	$x_n y_n$	$x_n x_n$
$\sum x$	$\sum y$	$\sum xy$	$\sum x^2$

Подставив в систему уравнений (2.6) численные значения $\sum x$, $\sum y$, $\sum x^2$ и $\sum xy$, найдите значения параметров линейного уравнения a и b.

2.3 Нахождение коэффициентов линейного уравнения с использованием ППП STATISTICA for WINDOWS

Проведем расчет параметров a и b линейного уравнения тренда вида (2.1) на компьютере с использованием подпрограммы *Множественная регрессия* пакета ***STATISTICA for WINDOWS*** по имеющемуся набору данных. Для этого выполним следующие действия с помощью кнопки *Статистика/Множественная регрессия.* Из двух вариантов расчета выберем *Быстрый*. После этого следует задать независимую *Var1 - (x)* и подчиненную *Var2 - (y)* переменные (рисунок 2.3).

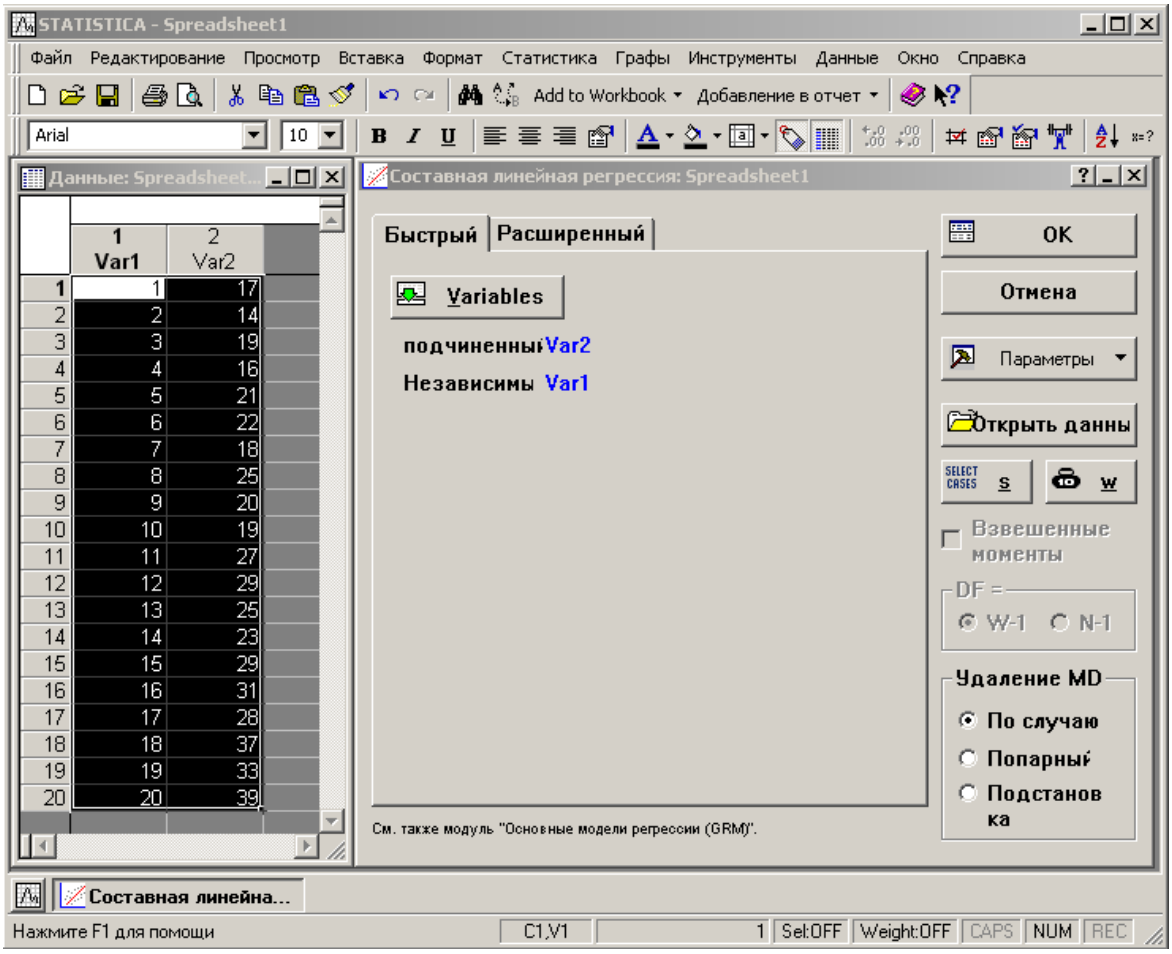


Рисунок 2.3 Окно множественной (составной) регрессии

Щелкните *ОК*, и после выполнения расчетов появится окно *Результаты составной регрессии* (рисунок 2.4), в котором выберете *Итог: результаты регрессии (Быстрый)*, получим окно (рисунок 2.5).

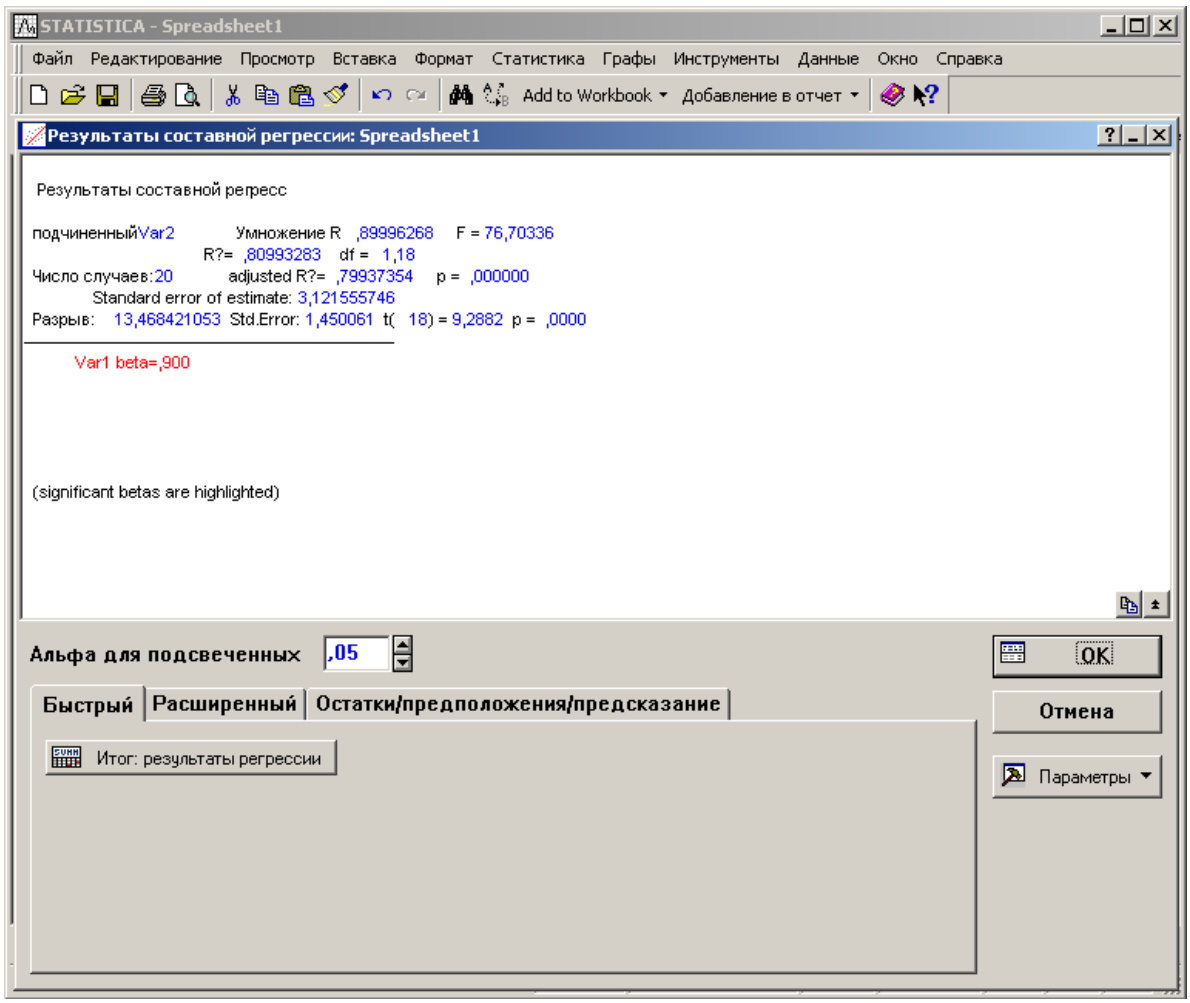


Рисунок 2.4 Окно *Результаты составной регрессии*

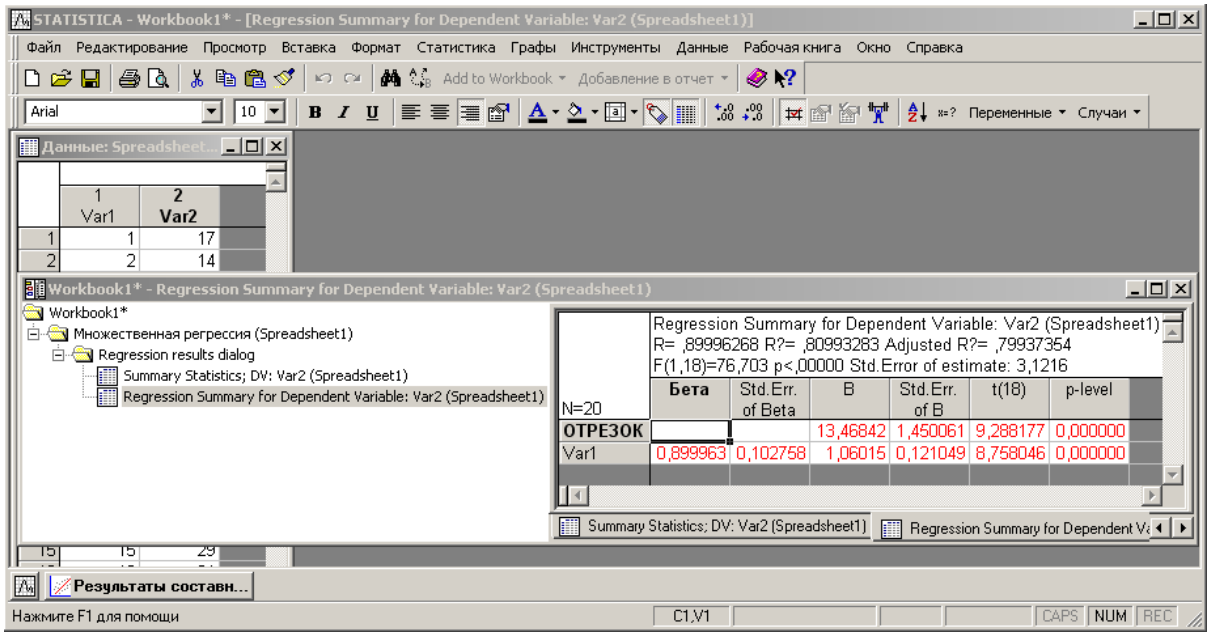


Рисунок 2.5 Основные результаты регрессионного анализа

В окне рисунок 2.5 коэффициенты уравнения линии регрессии показаны в столбце **B**: постоянная *a* (*ОТРЕЗОК*) и переменная *b* (*Var1*). Адекватность модели оценивается коэффициентом множественной корреляции *R*.

Для просмотра графиков выполним действия: *Графы/2D графы/Линейные вычерчивания (Переменные)* (рисунок 2.6).

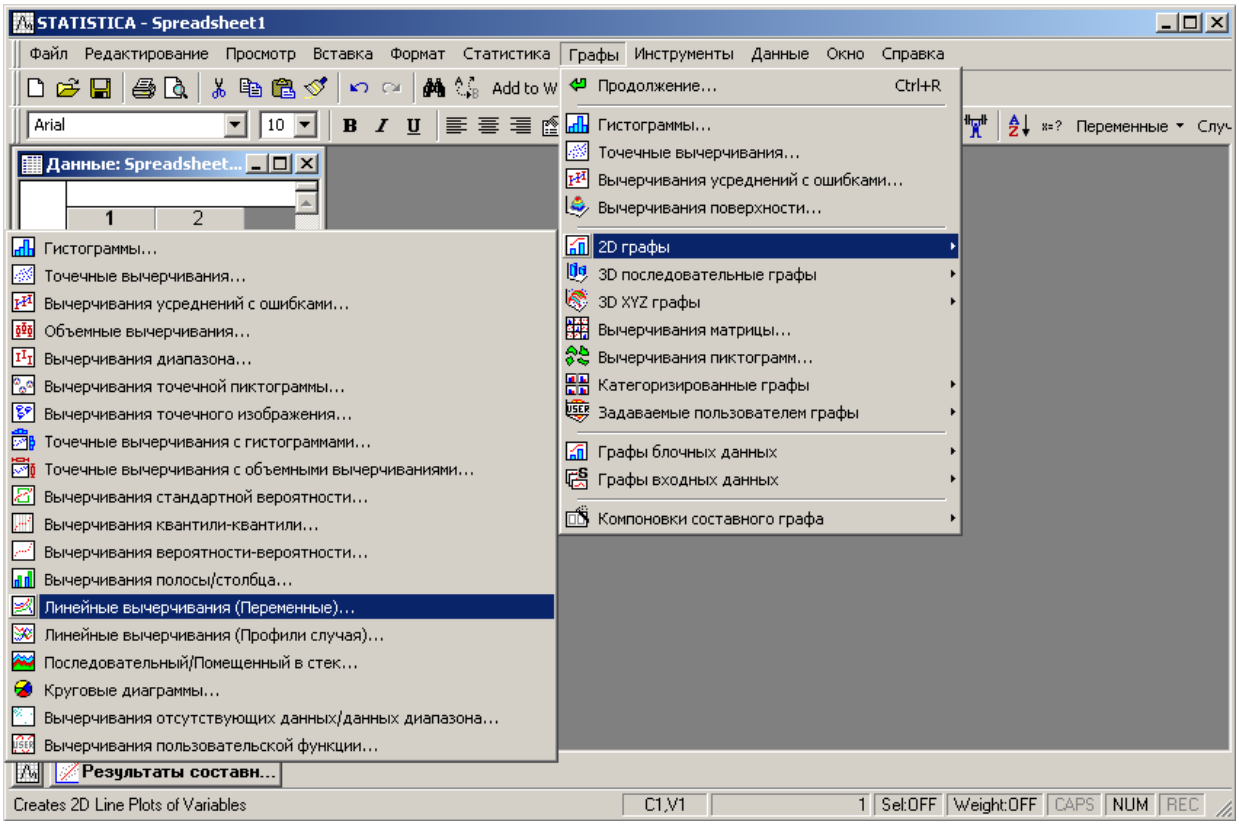


Рисунок 2.6 Выбор вида графиков

Щелкнув по кнопке *Линейные вычерчивания (Переменные)* получим окно *2D Line Plots* (рисунок 2.7). В этом окне установите раздел *Расширенный*, в котором выберете *Линейный* и *XY Trace*.

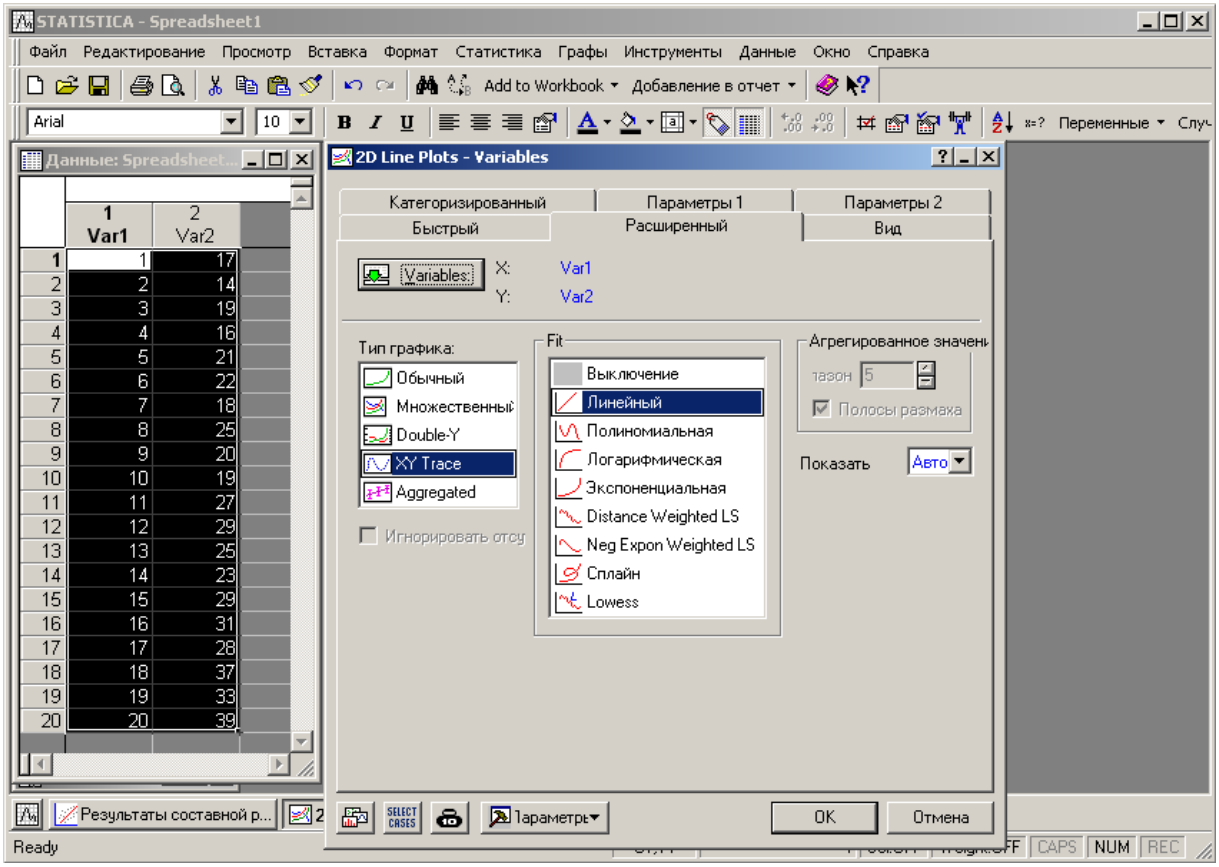


Рисунок 2.7 Окно выбора типа графика

Щелкните ОК, и после выполнения действий получим график линейной функции (рисунок 2.7). В этом окне показаны исходные данные (Case 1- Case *n*), линия регрессии и уравнение регрессии с численными значениями параметров уравнения *a* и *b*.

Сравните параметры уравнения *a* и *b*, полученные с помощью пакета ***STATISTICA for WINDOWS***, с аналогичными параметрами, рассчитанными вручную в разделе 2.1. Если Вы все выполнили правильно и пакет ***STATISTICA for WINDOWS*** также не ошибается, значения параметров уравнения *a* и *b* совпадут (не абсолютно точно – это объясняется различными способами расчета, а будут весьма близки).

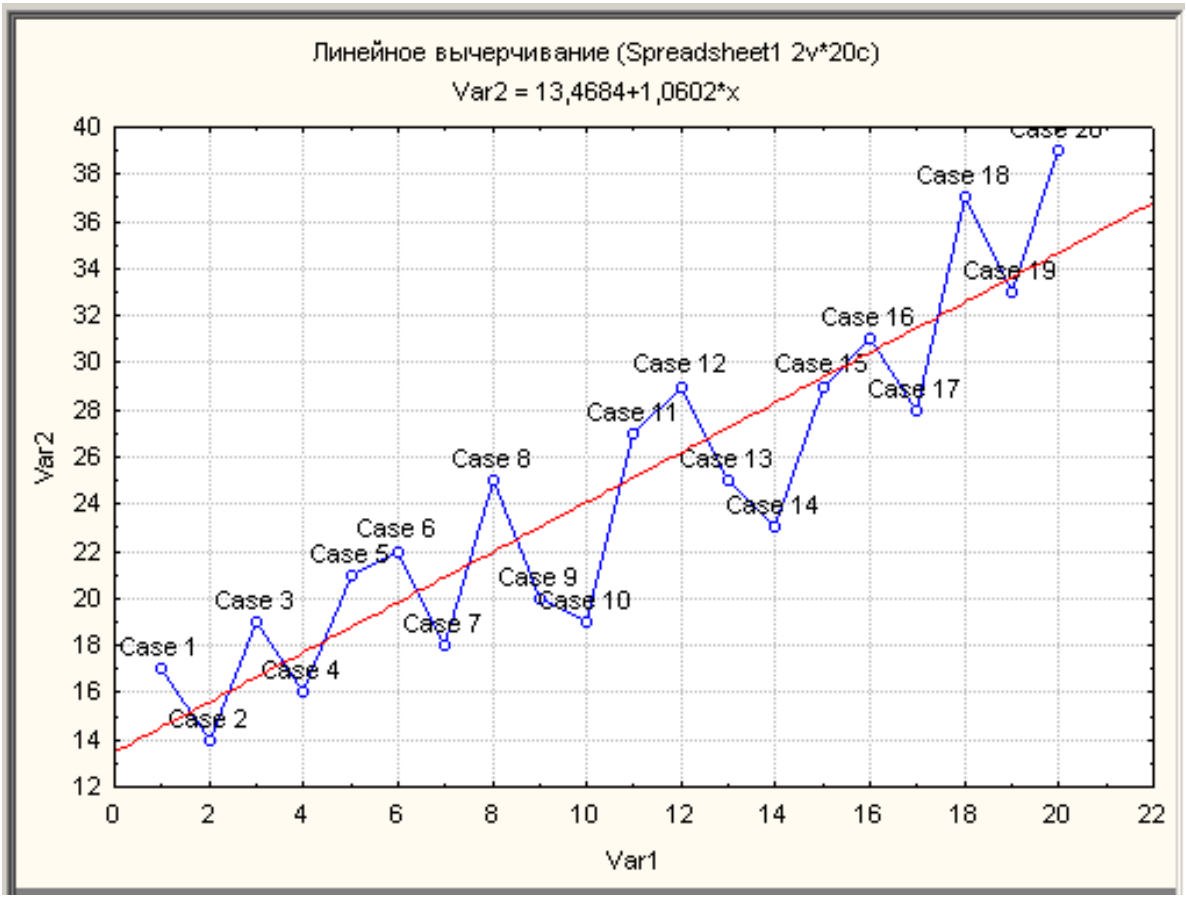


Рисунок 2.8 График линии регрессии и исходных данных

3 Построение многомерных регрессионных моделей

При выполнении экономических расчетов широко используются математические модели. Это функции спроса, предложения, потребления, инфляции и многие другие. Математические модели используются для прогнозирования, проведения экономического анализа, поддержки принятия решений и т.д. Математические модели могут быть линейными и нелинейными. Не затрагивая сейчас вопрос о виде зависимостей, заметим, что для расчетов нужны коэффициенты, т.е. численные параметры моделей. В экономике достаточно часто эти параметры определяются по результатам наблюдений статистическими методами, чаще всего посредством регрессионного анализа. С помощью анализа этих моделей можно определить условия, при которых деятельность Вашей фирмы будет оптимальной.

Для нахождения параметров многомерных регрессионных моделей используем тот же математический аппарат, который применялся для построения в предыдущем разделе для нахождения параметров линейного уравнения. В простейших случаях (как в предыдущем примере) пакет ***STATISTICA for WINDOWS*** позволяет сразу строить графики и выдает значения параметров уравнения. Но в более сложных случаях желательно иметь возможность задавать зависимости произвольного вида. Это могут быть двумерные функции *y(x)*, когда выходная переменная *y* зависит от одной входной переменной *x* или многомерные функции, когда выходная переменная *y* зависит от двух и более входных переменных $x_1, x_2, x_3, \ldots$

Рассмотрим ситуацию, когда на выходной параметр *y* влияют несколько входных параметров (обозначим их x_1 и x_2.). Какую модель имеет смысл построить: простую или сложную. Усложнение модели должно привести к повышению ее точности (статистической адекватности). Но, с другой стороны, за сложность приходится расплачиваться увеличением

числа наблюдений – количество наблюдений должно в 15-20 раз превышать число параметров модели. Поэтому желательно найти оптимум – достаточно точную и, в то же время, не слишком сложную модель.

Попробуем на основании исходного набора данных построить три модели:

1. Линейная двумерная:

$$y(x_1)=a_0+a_1x. \quad (3.1)$$

2. Квадратичная двумерная:

$$y(x_1) = a_0 + a_1x_1 + a_2x_1^2. \quad (3.2)$$

3. Квадратичная трехмерная:

$$y(x_1,x_2) = a_0 + a_1x_1 + a_2x_2 + a_3x_1^2 + a_4x_2^2 + a_5x_1x_2. \quad (3.3)$$

Сравним точность (адекватность) трех моделей по соответствующим коэффициентам множественной корреляции R (этот показатель уже рассматривали в предыдущей главе) .

$$R = \sqrt{1-\frac{S_e^2}{S_y^2}} \text{ , где} \quad (3.4)$$

S_e^2- квадрат среднеквадратичного отклонения ошибки (дисперсия отклонения);

S_y^2- квадрат среднеквадратичного отклонения случайной величины (дисперсия случайной величины).

Если коэффициенты множественной корреляции трех моделей практически не отличаются, то усложнение модели не имеет смысла. И, наоборот, при существенном увеличении значений коэффициентов множественной корреляции усложнение модели оправдано.

Указания к выполнению

- Расчет линейной двумерной модели (3.1) можно провести по методике *Множественная регрессия*, описанной в разделе построения тренда,

тем не менее, рассмотрим методы построения как линейной, так и нелинейных моделей (3.2) и (3.3) по единой методике.

- Составим таблицу произвольных исходных данных, например таблица 3.1. В этой базе данных роль y в моделях (3.1), (3.2), (3.3) выполняет $Var1, x_1 – Var2, x_2 – Var3$.

Таблица 3.1. База данных.

	1 Var1	2 Var2	3 Var3
1	24	23	42
2	67	34	33
3	35	41	25
4	89	52	12
5	121	47	23
6	31	69	12
7	57	50	8
8	34	84	10
9	84	72	14
10	68	93	9
11	53	88	7
12	47	95	5

3.1 Построение линейной двумерной модели

$$y(x_1) = a_0 + a_1 x_1.$$

- Запустите окно *Дополнительные линейные/нелинейные модели* и выберите *Нелинейный подсчет* (рисунок 3.1).

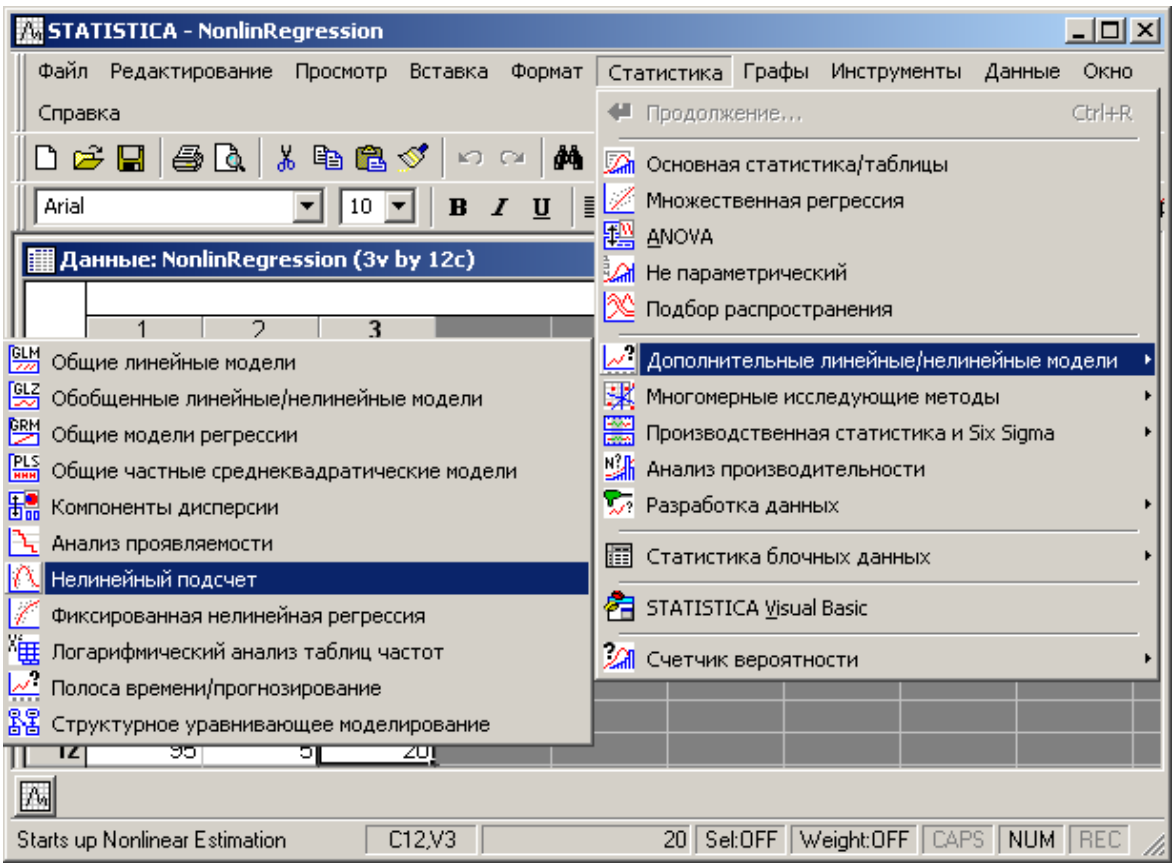


Рисунок 3.1 Окно *Дополнительные линейные/нелинейные модели*

- Щелкните на кнопку *Нелинейный подсчет,* в результате появится окно выбора процедур (рисунок 3.2), в котором выберете *Задаваемая пользователем регрессия.*

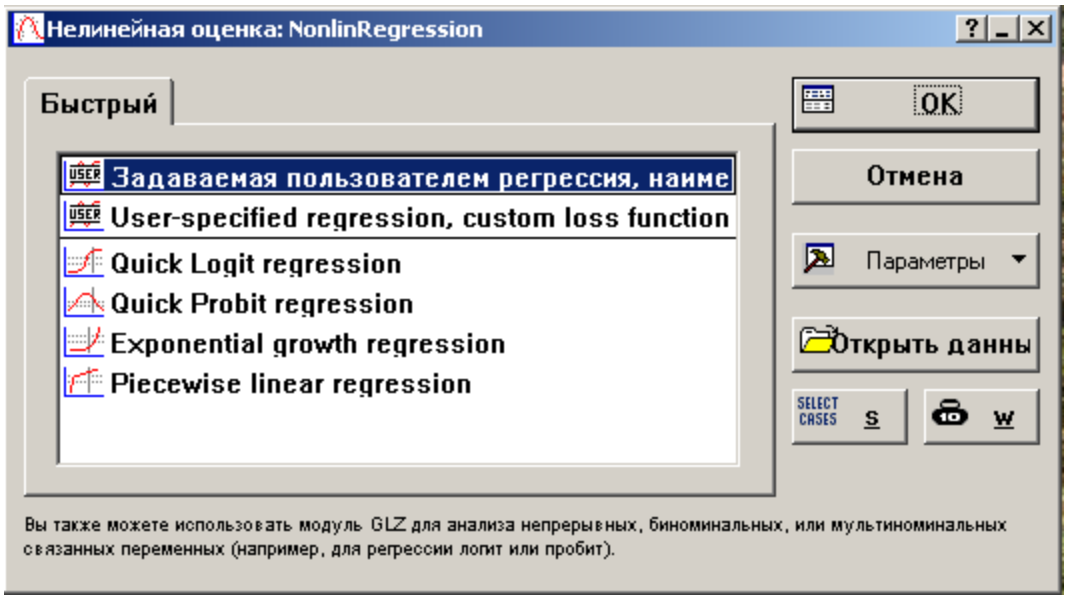


Рисунок 3.2 Окно выбора процедур

- Щелкните на кнопку *Задаваемая пользователем регрессия,* в результате появится окно *Функция для оценки* (рисунок 3.3), в котором пока отсутствует задаваемая функция.

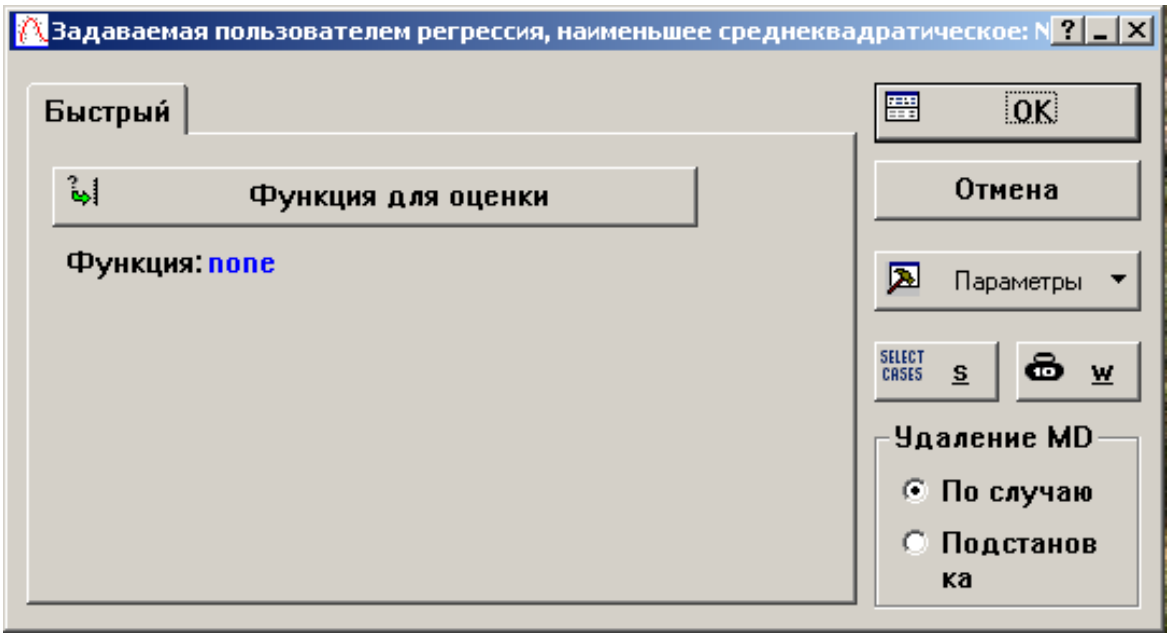


Рисунок 3.3 Окно *Функция для оценки* (без заданной функцией)

- Щелкните на кнопку *Функция для оценки,* в результате появится окно *Оцениваемая функция* (рисунок 3.4), в котором задайте функцию v1 =a0+a1*v2, соответствующую линейному уравнению (3.1).

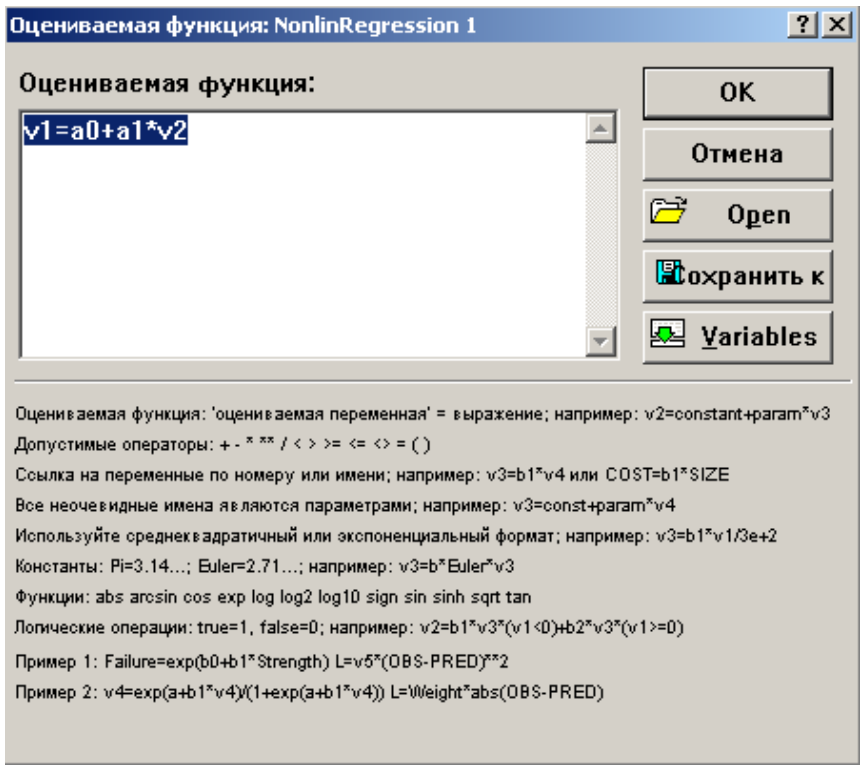


Рисунок 3.4 Окно *Оцениваемая функция*

- После подтверждения оцениваемой функции, вернемся в предыдущее окно *Функция для оценки,* в котором эта функция уже задана (рисунок 3.5).

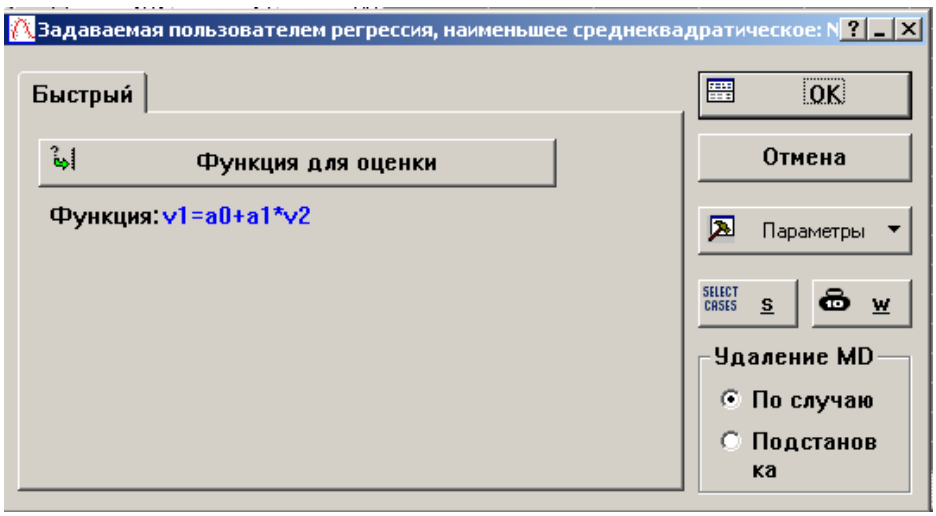


Рисунок 3.5 Окно *Функция для оценки* (с заданной функцией)

- После подтверждения в этом окне оцениваемой функции, получим окно задания способов расчета модели (рисунок 3.6), в котором в разделе *Расширенный* зададим алгоритм поиска параметров модели *Метод Оценки: Gauss-Newton*.

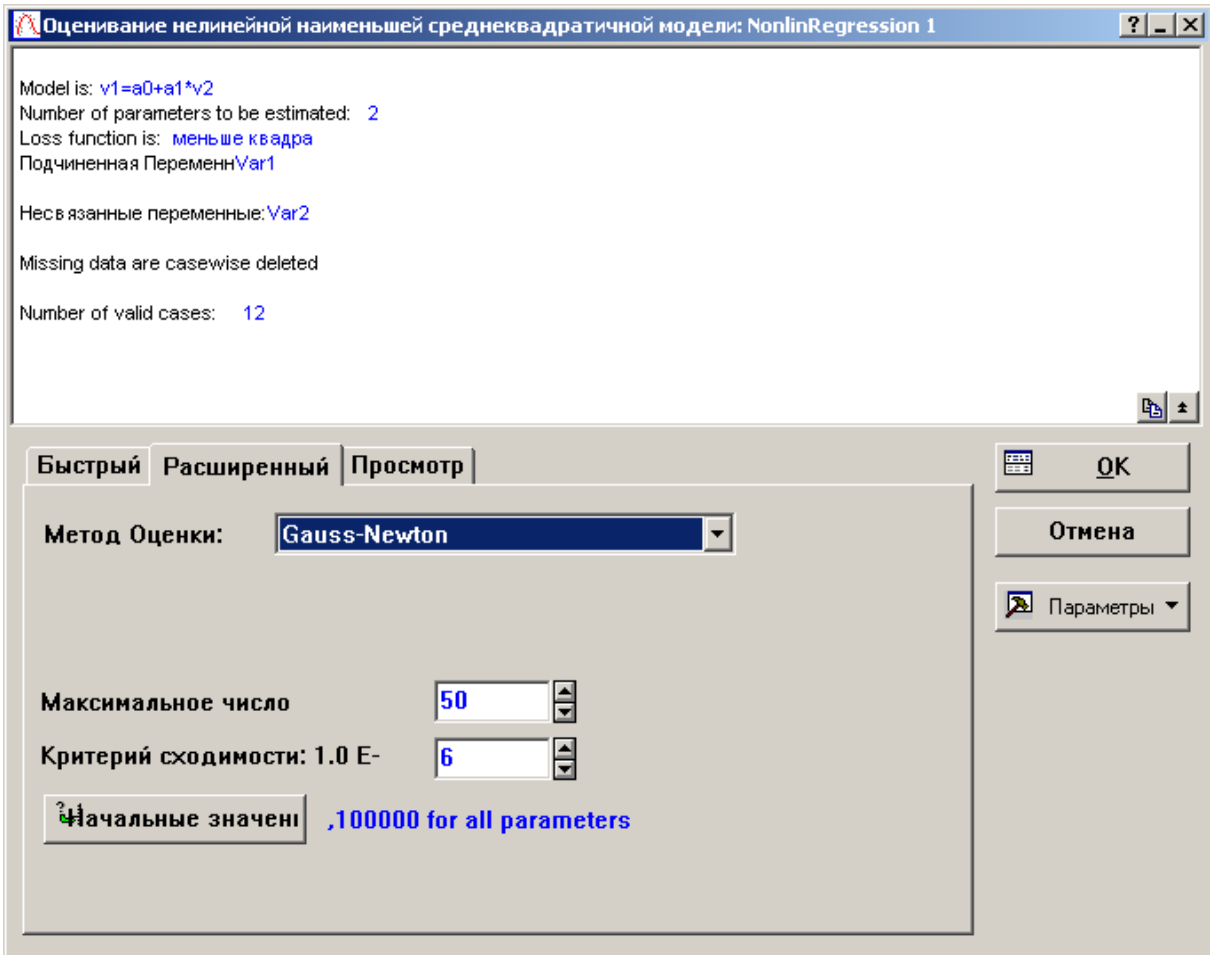


Рисунок 3.6 Окно задания способов расчета модели

- После подтверждения алгоритма поиска параметров модели и всех условий расчета в окне (рисунок 3.6) получим окно с результатами расчета (рисунок 3.7). Главным показателем адекватности модели является коэффициент множественной корреляции R. С учетом того, что этот расчет выполнен для первой модели (3.1) из трех, будем рассматривать этот коэффициент как R_1. В данном примере R_1 =0,0432531.

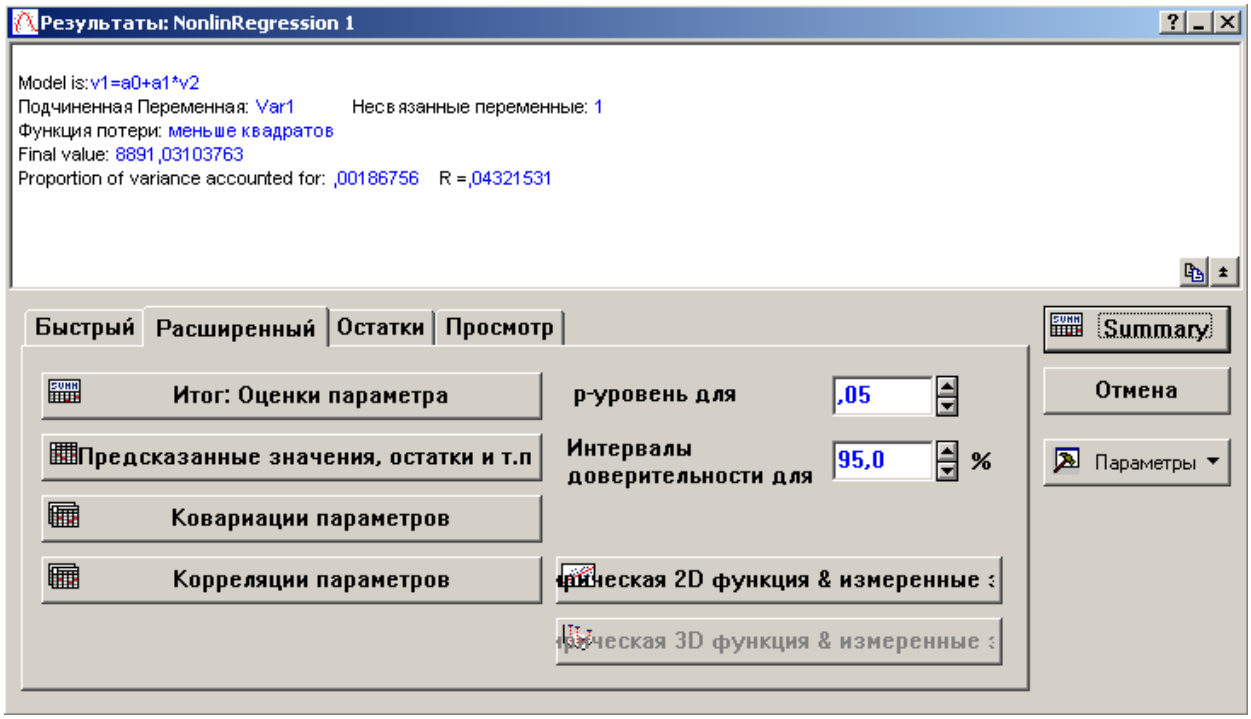
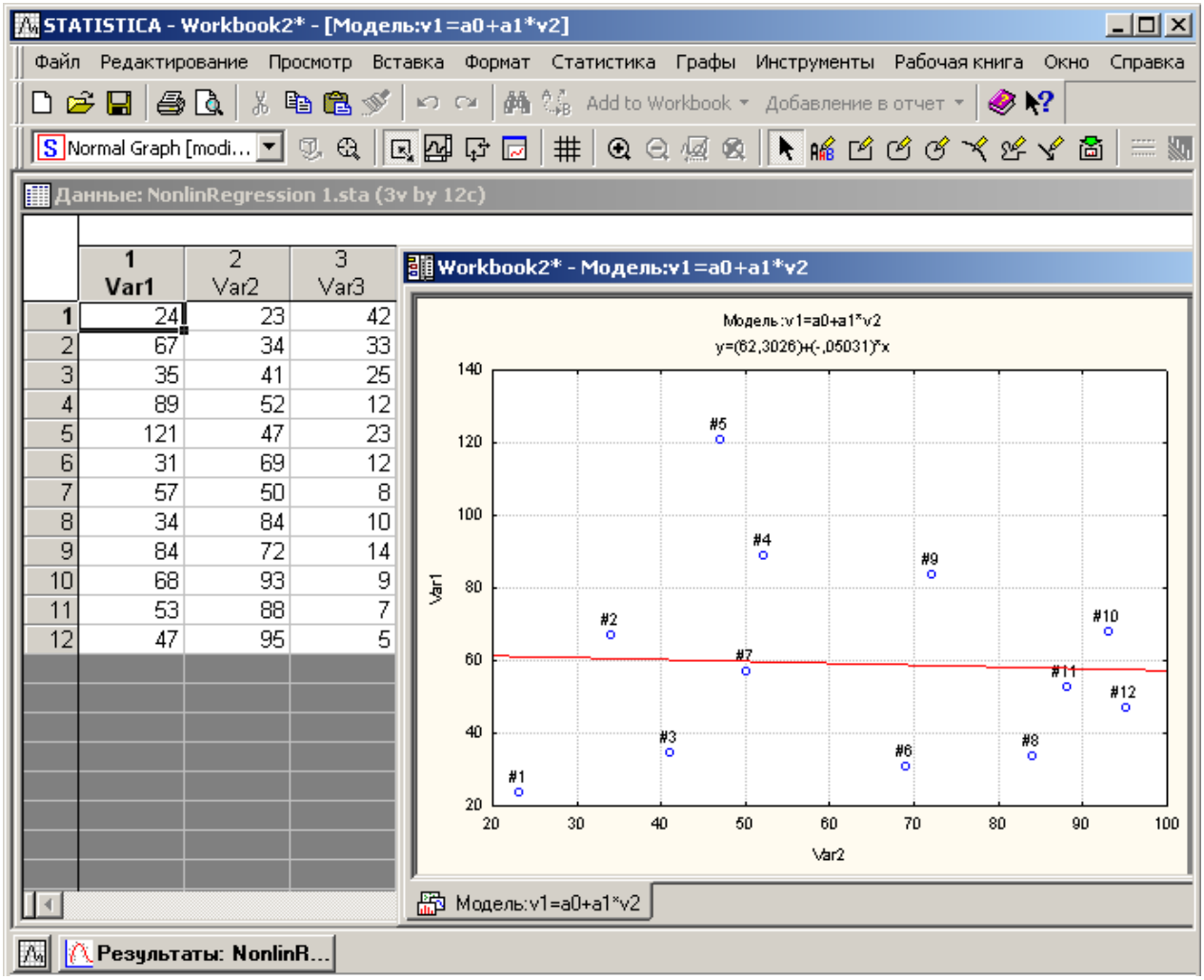


Рисунок 3.7 Окно результатов расчета линейной модели (3.1).

- Для просмотра графика щелкнем по кнопке 2D функция в окне (рисунок 3.7). В результате получим линейный график (рисунок 3.8). Помимо самого графического изображения в окне показан аналитический вид функции и ее численные параметры.



	1 Var1	2 Var2	3 Var3
1	24	23	42
2	67	34	33
3	35	41	25
4	89	52	12
5	121	47	23
6	31	69	12
7	57	50	8
8	34	84	10
9	84	72	14
10	68	93	9
11	53	88	7
12	47	95	5

Рисунок 3.8 Линейная функция и ее график

3.2 Построение квадратичной двумерной модели

$$y(x_1) = a_0 + a_1x_1 + a_2x_1^2.$$

- В окне *Функция для оценки* зададим функцию v1=a0+a1*v2+a2*v2*v2 (рисунок 3.9).

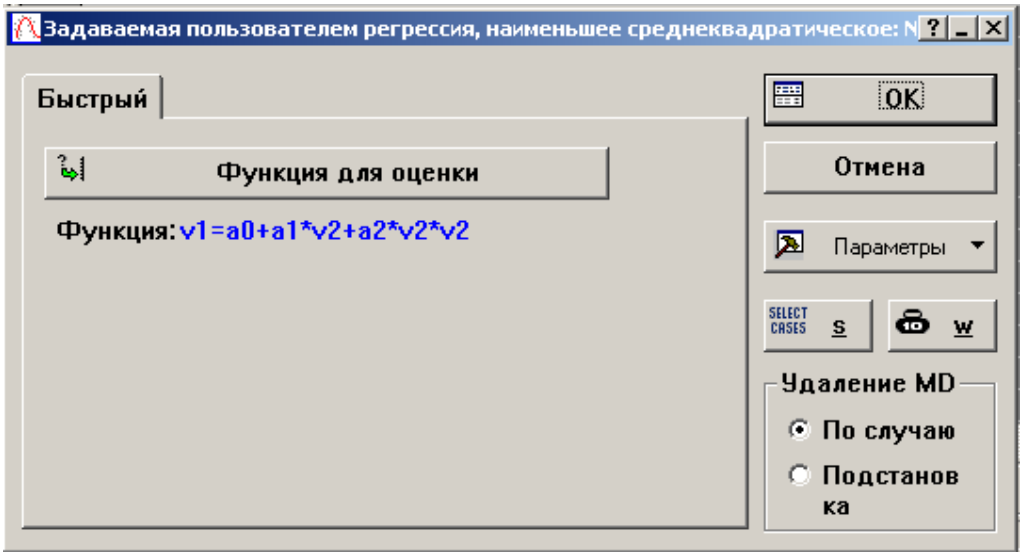


Рисунок 3.9 Окно *Функция для оценки* квадратичной двумерной модели

- После проведения расчетов аналогично предшествующей модели получим окно (рисунок 3.10) с результатами расчета квадратичной двумерной модели. С учетом того, что этот расчет выполнен для второй модели (3.2) из трех, будем рассматривать полученный коэффициент множественной корреляции R как R_2. В данном примере $R_2 = 0{,}42236071$.

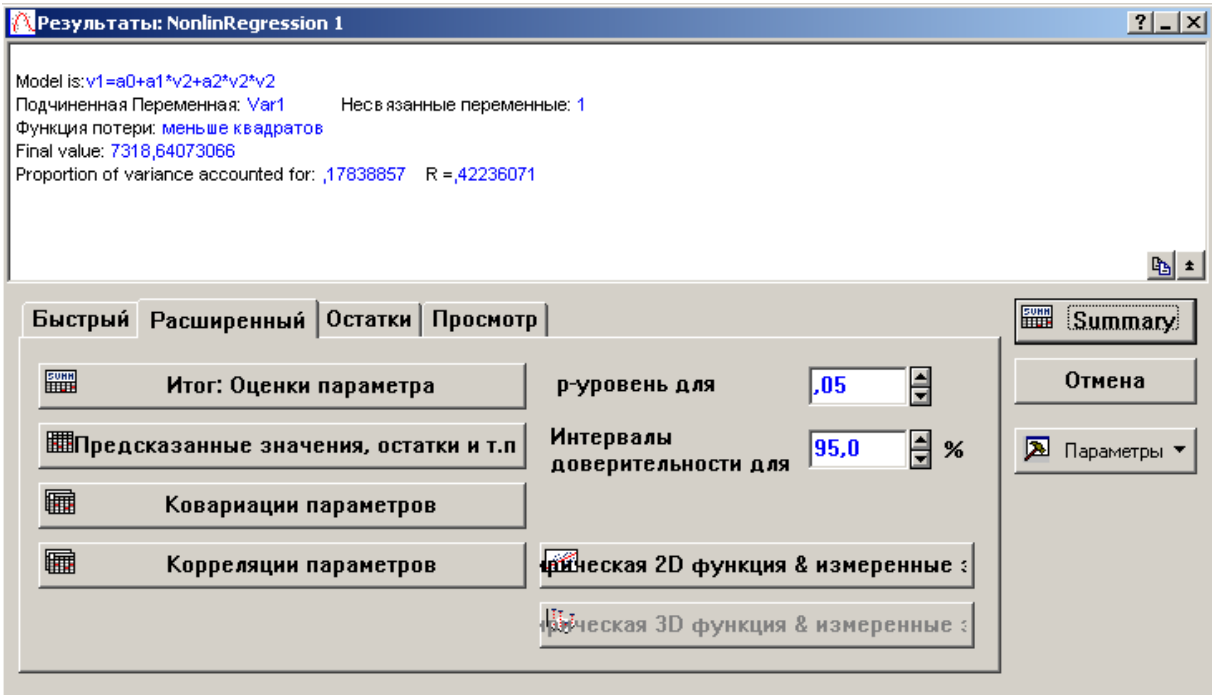
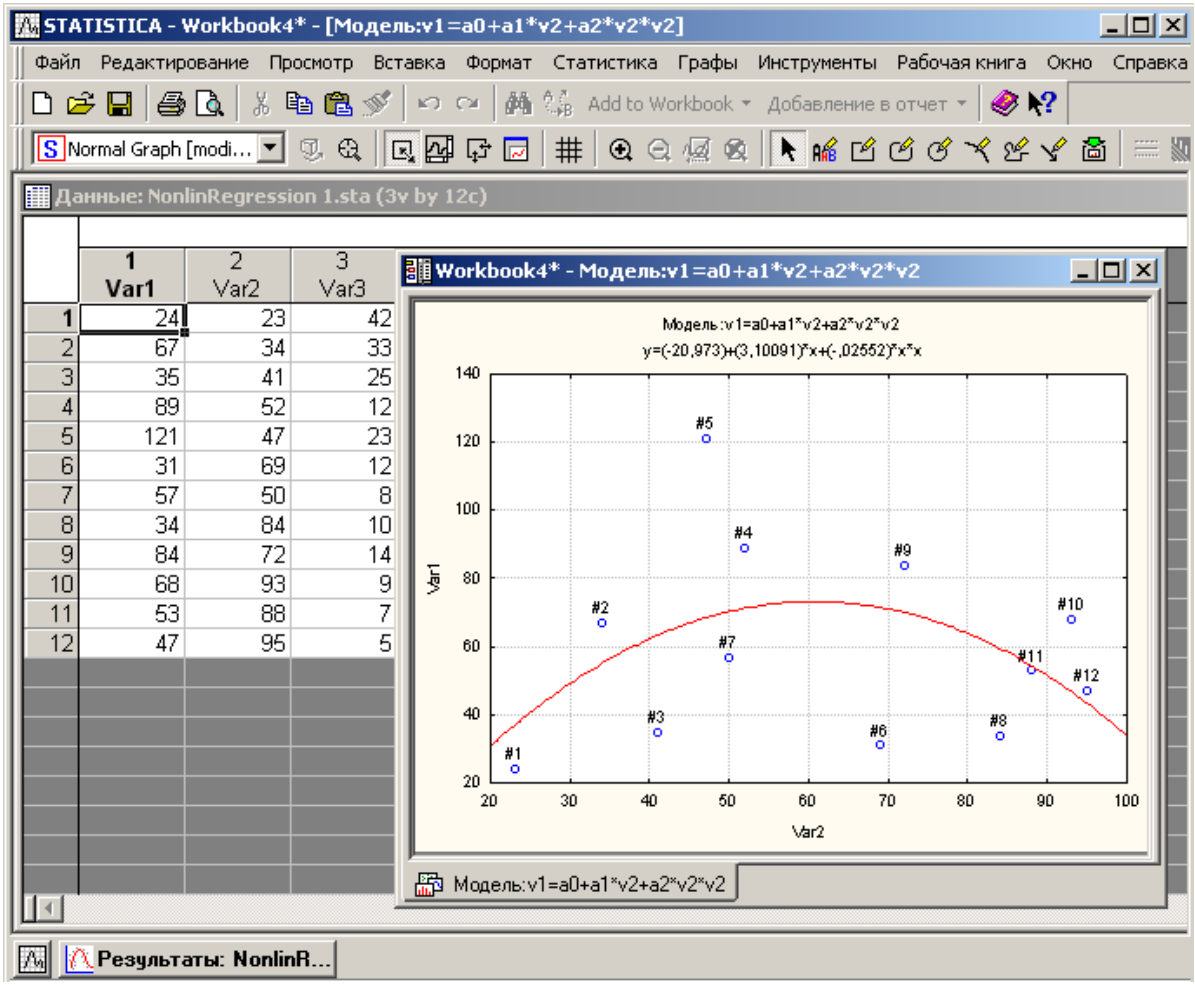


Рисунок 3.10 Окно результатов расчета квадратичной двумерной модели (3.2)

- Щелкните по кнопке 2D функция в окне (рисунок 3.10). В результате получим нелинейный график (рисунок 3.11).



	1 Var1	2 Var2	3 Var3
1	24	23	42
2	67	34	33
3	35	41	25
4	89	52	12
5	121	47	23
6	31	69	12
7	57	50	8
8	34	84	10
9	84	72	14
10	68	93	9
11	53	88	7
12	47	95	5

Рисунок 3.11 Нелинейная функция и ее график

3.3 Построение квадратичной трехмерной модели

$$y(x_1,x_2) = a_0 + a_1x_1 + a_2x_2 + a_3x_1^2 + a_4x_2^2 + a_5x_1x_2.$$

- В окне *Функция для оценки* зададим функцию v1=a0+a1*v2+a2*v3+a3*v2*v2+a4*v3*v3+a5*v2*v3 (рисунок 3.12).

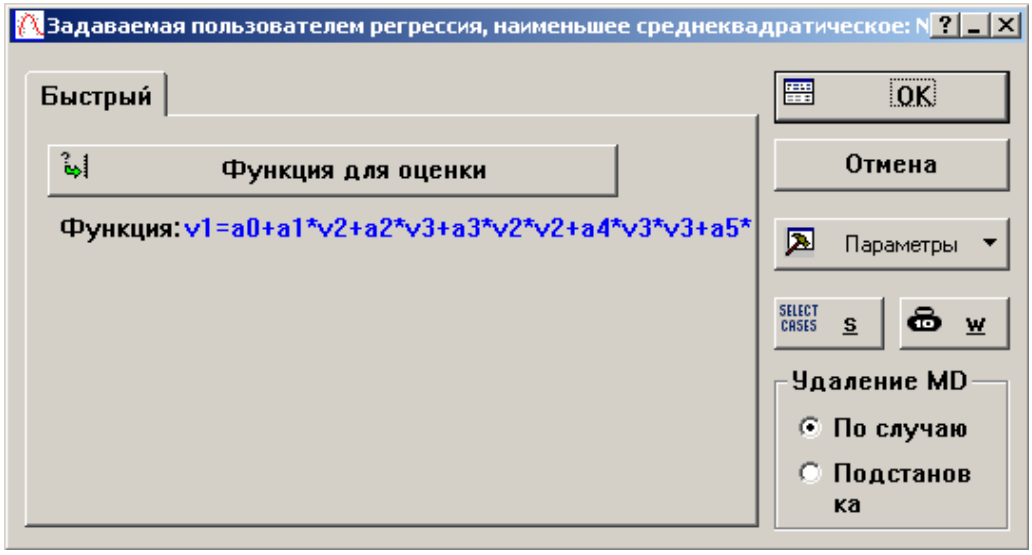


Рисунок 3.12 Окно *Функция для оценки* квадратичной трехмерной модели

- Подтвердив заданную функцию и алгоритм расчета, получим окно результатов расчета квадратичной трехмерной модели (рисунок 3.13). С учетом того, что этот расчет выполнен для третьей модели (3.3) из трех, будем рассматривать этот коэффициент как R_3. В данном примере $R_3 = 0{,}56614999$.

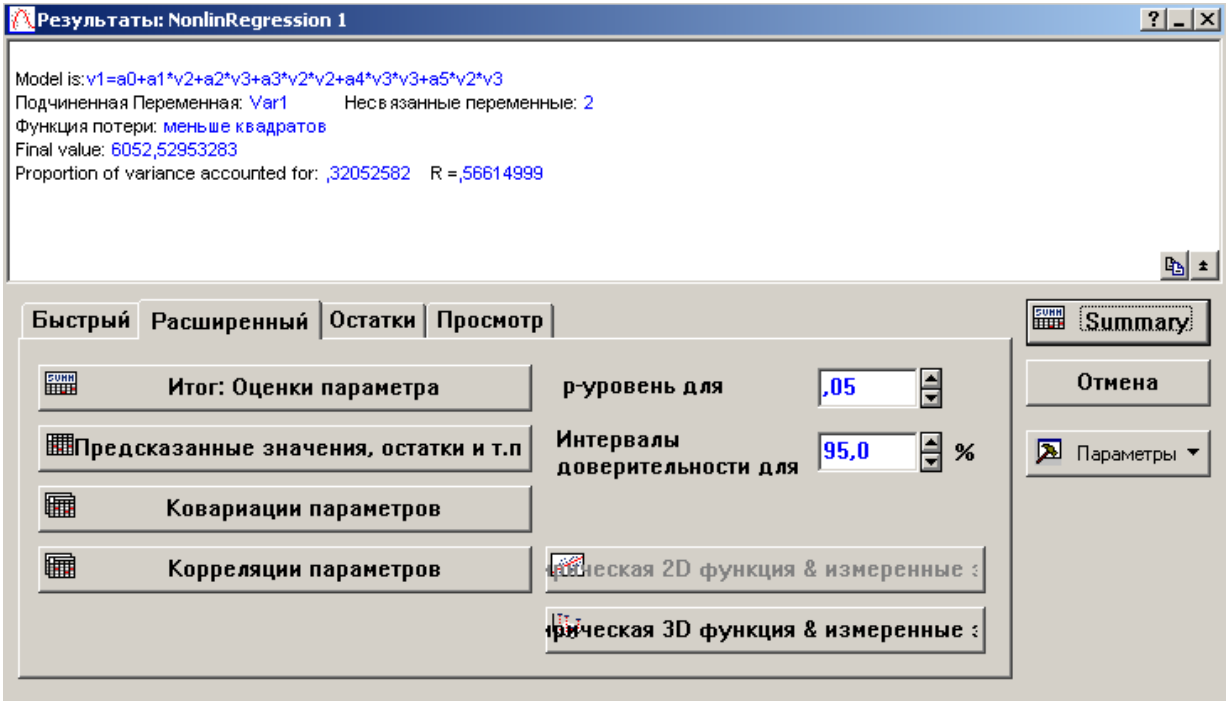


Рисунок 3.13 Окно результатов расчета квадратичной трехмерной модели (3.3)

- Для просмотра поверхности отклика щелкните по кнопке 3D функция. Получим окно с желаемой поверхностью (рисунок 3.14).

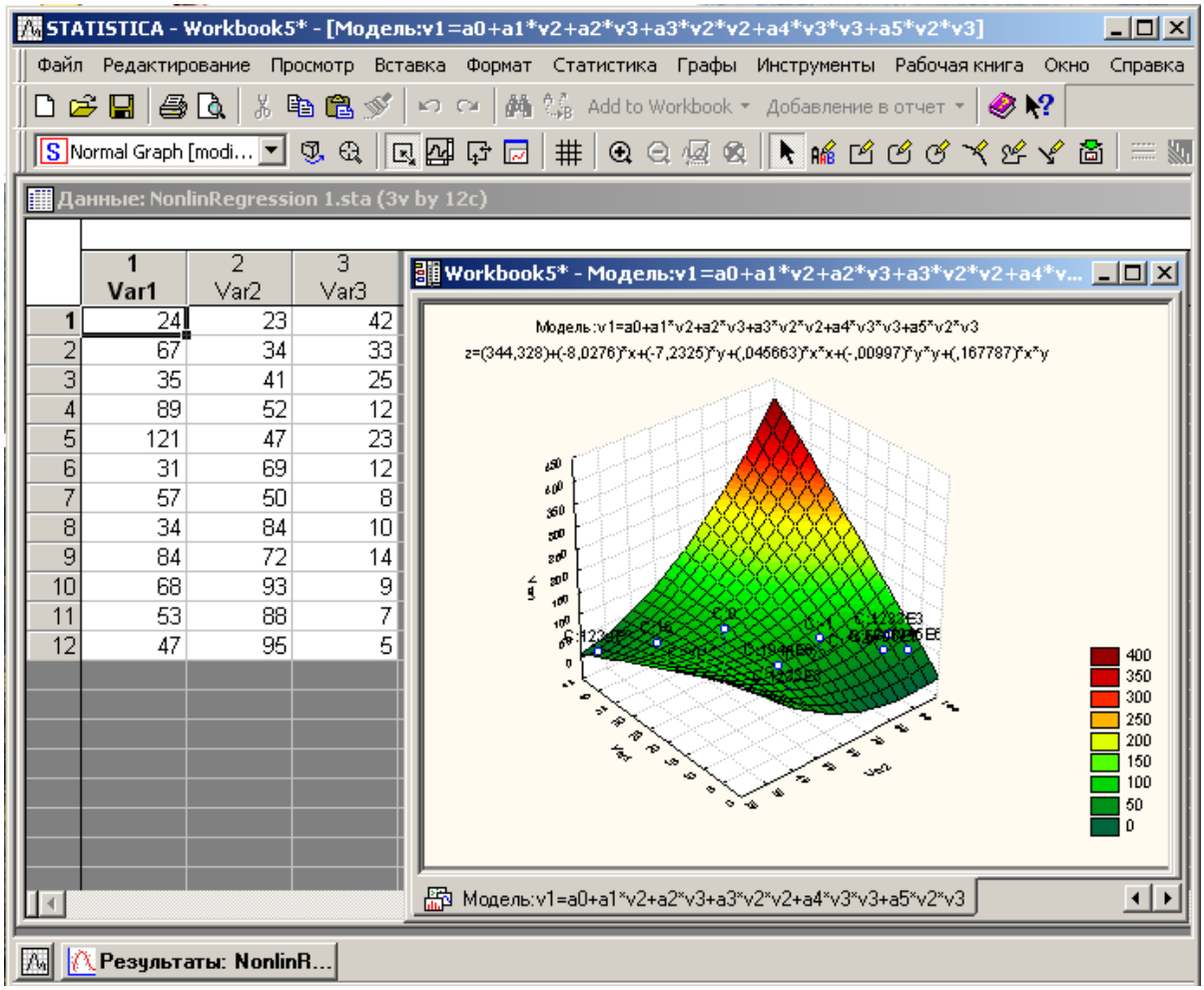


Рисунок 3.14 Нелинейная функция, и ее поверхность отклика

3.4 Определение оптимальной структуры модели.

Сравните коэффициенты множественной корреляции трех моделей: R_1 – линейной (3.1), R_2 – квадратичной двумерной (3.2) и R_3 – квадратичной трехмерной (3.3). Оцените целесообразность усложнения моделей. Если значения коэффициентов множественной корреляции существенно возрастают при добавлении новых членов в структуру модели, тогда такое усложнение моделей оправдано. В противном случае целесообразно остановиться на самой простой структуре модели.

4 Оценка чувствительности Вашей фирмы к …

4.1 Исследование анализа чувствительности

Обладание построенной моделью предоставляет возможность провести интересное исследование на тему: ***что будет, если?*** Такие вопросы часто возникают перед каждым менеджером, так как он вынужден действовать в непрерывно изменяющемся мире. Насколько изменится цена изделия при повышении стоимости энергоносителей, как скажется на зарплате рост производительности труда и т.д. Ответы на подобные вопросы пытаются найти специалисты с помощью анализа различных моделей, как аналитических, так и статистических. Если Вы построили двумерную или трехмерную статистическую модель, у Вас имеется возможность провести анализ чувствительности, причем не только ретроспективно, но и предположительно ***что будет, если?***

Для выполнения анализа чувствительности нужно дважды рассчитать значение выходного параметра: при номинальном значении входного параметра и при увеличенном (уменьшенном) значении на 5% или 10%. Необходимо иметь в виду, что модели обладают достоверностью при небольшом изменении аргументов. Если входной параметр изменить в несколько раз, то такой расчет не будет достоверен.

4.2 Указания к выполнению

- Если имеется двумерная линейная модель $y(x_1) = a_0 + a_1x_1$ (3.1) или квадратичная двумерная $y(x_1) = a_0 + a_1x_1 + a_2x_1^2$ (3.2), то, подставив значение x_{1i}, соответствующее его номинальной величине в исследуемом диапазоне, получим значение $y(x_{1i})$. Затем произведем расчет выходной переменной для увеличенного на 5% значения

входной переменной $x_{1i.}$ Получим значение $y(1{,}05\ x_{1i})$, и сравним $y(1{,}05\ x_{1i})$ и $y(x_{1i})$. На основании этого можно сделать вывод, что при увеличении входной переменной x_1 на 5%, выходная переменная $y(x_1)$ изменится на … %.

- При анализе линейной модели чувствительность будет неизменной при любом значении входной переменной x_1, так как чувствительность оценивается производной функции. Но если исследуется нелинейная функция, то чувствительность выходной переменной $y(x_1)$ будет меняться в зависимости от значения входной переменной x_1 (рисунок 4.1). Поэтому нужно обоснованно выбрать положение значения входной переменной для оценки чувствительности.

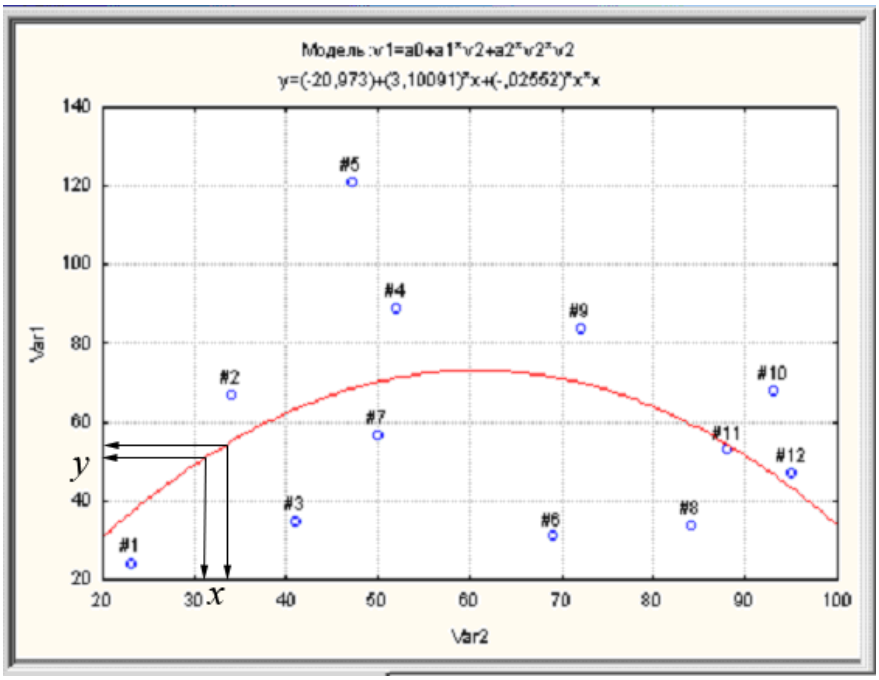


Рисунок 4.1 Чувствительность меняется в зависимости от положения точки

- Если имеется квадратичная трехмерная модель $y(x_1,x_2) = a_0 + a_1x_1 + a_2x_2 + a_3x_1^2 + a_4x_2^2 + a_5x_1x_2$ (3.3), то ситуация еще сложнее. Выходная переменная зависит от двух входных переменных (рисунок 4.2). Мы можем исследовать чувствительность одновременно только от одной входной переменной, заморозив (задав неизменное значение) другую входную переменную.

- Зададимся какими-либо значениями x_{1i} и x_{2j}, соответствующее их номинальным величинам в исследуемом диапазоне, получим значение $y(x_{1i},x_{2j})$. Если нужно исследовать чувствительность выходной переменной к изменению x_1, то проведем расчет выходной переменной для увеличенной на 5% значения входной переменной x_{1i} при неизменном значении x_{2j}. Получим значение $y(1,05\ x_{1i},x_{2j})$, и сравним $y(1,05\ x_{1i},x_{2j})$ и $y(x_{1i},x_{2j})$. На основании этого можно сделать вывод, что при увеличении входной переменной x_1 на 5% выходная переменная $y(x_1,x_2)$ изменится на … %.

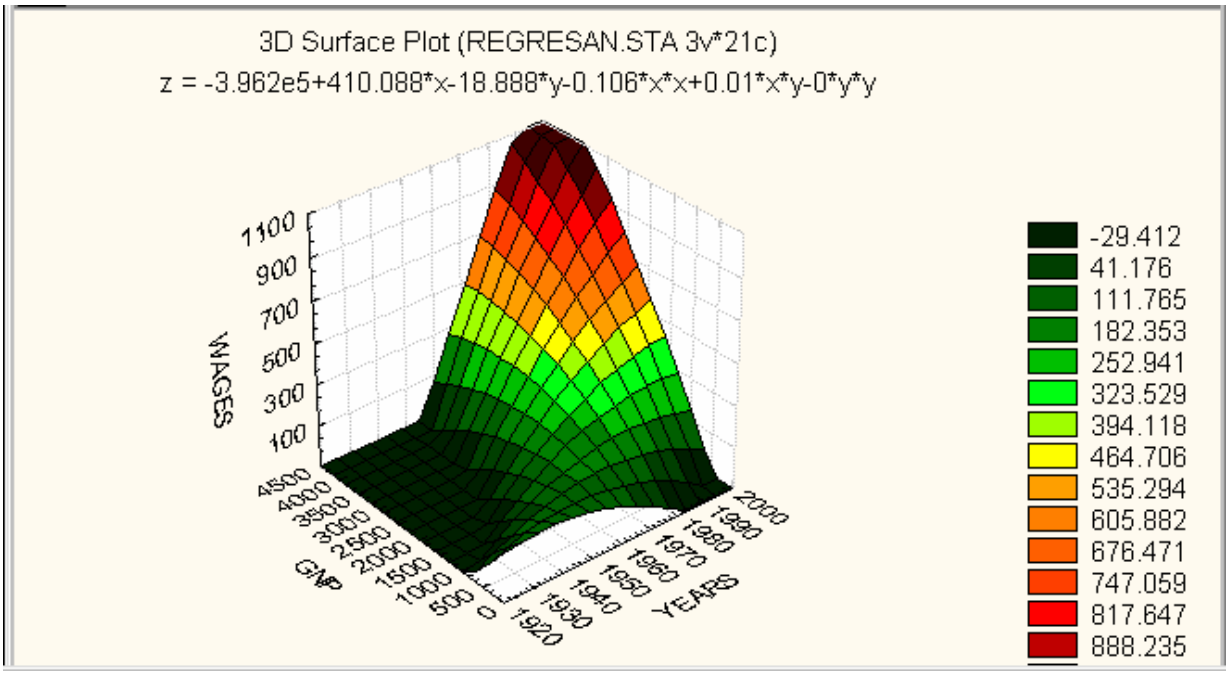


Рисунок 4.2 Поверхность отклика трехмерной модели

- При необходимости исследования чувствительности выходного параметра к изменению x_2, проделаем расчеты аналогично предыдущим, но при этом будем увеличивать x_2 при неизменном x_1.
- Много это или мало, т.е. насколько высока чувствительность выходной переменной к изменению входных переменных, можно судить в каждом конкретном случае отдельно, исходя из существа исследуемых зависимостей.

5 Куда движется Ваша фирма?

5.1 Выделение тренда из экспериментальных данных

В окружающей нас внешней среде имеется огромное количество явлений и объектов, изменяющихся в пространстве и во времени, и любой набор данных, состоящий из упорядоченных по этим координатам измерений, может рассматриваться как динамический ряд. Жизнь идет, со временем меняются показатели работы фирмы. Чтобы выявить тенденции их изменений, выделим из набора данных направленность (тренд), т.е. постепенное изменение за длительное время или на больших расстояниях.

Наиболее эффективным методом описания тренда является подбор некоторого аналитического выражения, как правило, полинома невысокой степени. Более высокая степень полинома нуждается в смысловом обосновании, поэтому для начала ограничим поиски 1-й степенью полинома.

Представим тренд в виде прямой:

$$y = a + bt, \qquad (5.1)$$

где t - время, y - выходной параметр.

Фактически уравнение (5.1) аналогично линейному уравнению (3.1), в котором вместо входной переменной t использовалась входная переменная x. Т.е. для построения тренда можно воспользоваться процедурами построения регрессионных моделей, описанных в главе 3. Если для описания тренда нужно применять более сложную зависимость, например, для получения более высокой степени статистической адекватности - коэффициента множественной корреляции R (3.4), используйте процедуры построения моделей произвольного, задаваемого пользователем вида.

После построения модели тренда можно подставить в эту модель нужное будущее время, и составить, таким образом, прогноз развития фирмы. При этом необходимо помнить, что задаваемое будущее время не

должно намного отличаться от исследованного диапазона, в противном случае точность прогноза будет снижаться.

5.2 Указания к выполнению

Для составления прогноза развития фирмы и оценки его погрешности нужно иметь набор данных за определенный период. Чтобы можно было достоверно определить, насколько может быть точен прогноз, выполним следующие действия:

- Построим линейную модель вида $y = a + bt$ (или нелинейную полиномиальную двумерную модель), не за весь период исследования, а несколько меньший. Определим параметры модели a и b.
- Вычислим прогнозные значения исследуемого параметра фирмы y_n. Для этого подставим в полученную модель значение y_r, соответствующее конечному времени периода.
- Сравним вычисленное прогнозное значение y_n с реальным y_r и определим погрешность прогноза Δ по формуле $\Delta = \frac{y_r - y_n}{y_r} 100\%$.

- Например, на рисунке 5.1 показано изменение валового национального продукта США с 1930 по 1990 годы. Можно по данным за 1930 по 1980 годы построить модель тренда (рисунок 5.2). Затем в эту модель нужно подставить 1990 год и рассчитать прогнозное значение GNP_n за 1990 год. После этого сравните прогнозное значение GNP_n с реальным GNP_r и рассчитайте погрешность прогноза.

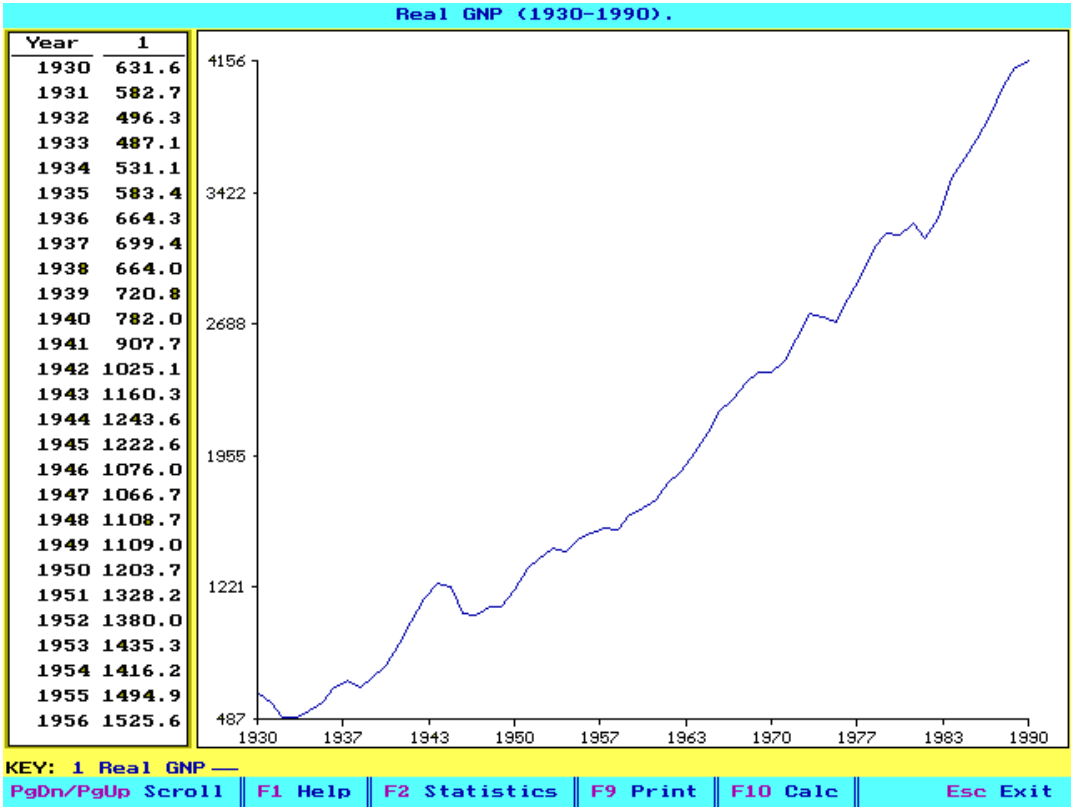


Рисунок 5.1 Изменение валового национального продукта США с 1930 по 1990 годы

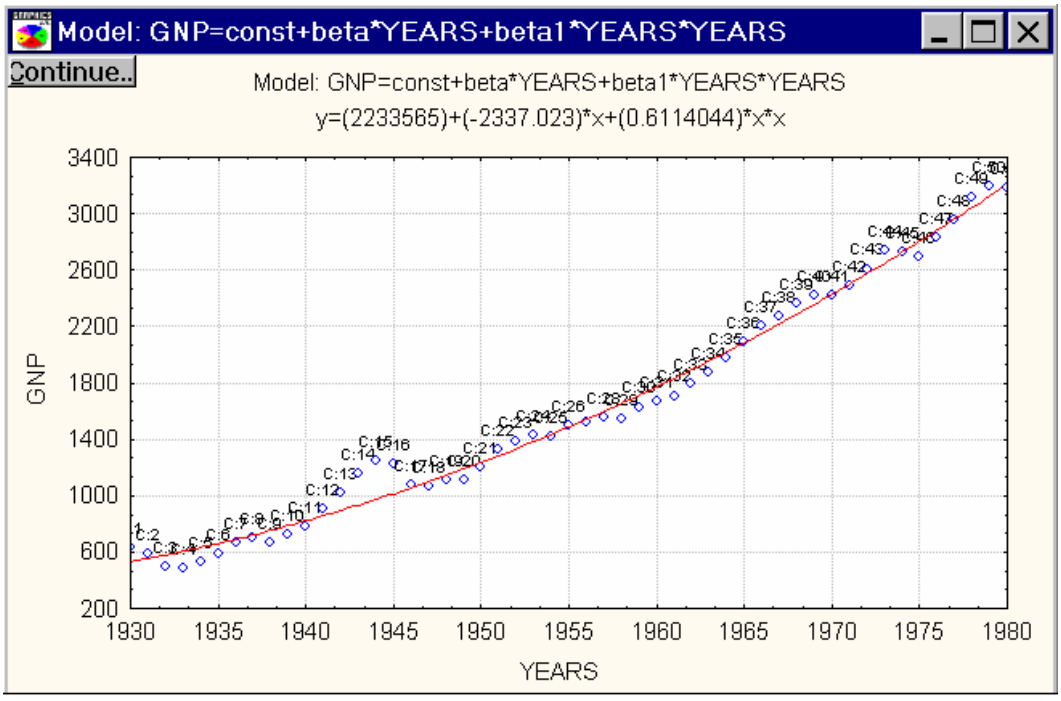


Рисунок 5.2 График тренда валового национального продукта США

- Если погрешность прогноза по прошедшему периоду достаточно низкая, Вы можете использовать полученную модель для составления прогноза на будущее.

6 От каких факторов зависит успех Вашей фирмы?

6.1 Основы многомерного факторного анализа

Большинство объектов, в том числе и Ваша фирма, как правило, подвержены действию сразу множества факторов. Реально затруднительно контролировать и управлять всеми ими. Т.е. объекты являются частично управляемыми и наблюдаемыми. Следовательно, при разработке методов управления экономическими объектами приходится решать, какие факторы являются самыми важными. Эти факторы необходимо измерять и далее включать их в математические модели, а остальными можно, в крайнем случае, пренебречь.

Задача определения степени влияния факторов на выходной параметр объекта управления, (например, прибыль фирмы) решается с помощью факторного анализа по доле дисперсии выходного параметра. Полная дисперсия $D_y = S_y^2$ выходной переменной y может быть разбита на три основных компонента:

1. *Общую дисперсию*, которую можно определить как часть дисперсии переменной y, появляющейся под действием k факторов.

2. *Специфическую дисперсию*, представляющую собой часть, которая не связана с исследуемыми переменными.

3. Дисперсию, обусловленную ошибкой измерения. Она является случайной, вызванной ошибками в процессе выборки, отклонениями от условий эксперимента и т.д.

$$S^2{}_y = S^2{}_{y_1} + S^2{}_{y_2} + \ldots + S^2{}_{y_k} + S^2{}_{y_\lambda} + S^2{}_{y_\delta}, \quad (6.1)$$

где $\sum_{i=1}^{k} S_{y_i}^2$ – общая дисперсия;

$S^2{}_{y_\lambda}$ – специфическая дисперсия;

$S^2{}_{y_\delta}$ – дисперсия, обусловленная ошибкой измерения.

Разделив обе части уравнения (6.1) на $S^2{}_y$, получим:

$$1 = \frac{S^2{}_{y_1}}{S^2{}_y} + \frac{S^2{}_{y_2}}{S^2{}_y} + \ldots + \frac{S^2{}_{y_k}}{S^2{}_y} + \frac{S^2{}_{y_\lambda}}{S^2{}_y} + \frac{S^2{}_{y_\delta}}{S^2{}_y}. \qquad (6.2)$$

Обозначим отношения в уравнении (6.2) соответственно, как $F^2_{y_i}$, Λ^2_y и Δ^2_y. Таким образом, полная (нормированная) дисперсия переменной y равна 1, а все составляющие дисперсии в правой части уравнения представляют доли в полной дисперсии:

$$1 = F^2_{y_1} + F^2_{y_2} + \ldots + F^2_{yk} + \Lambda^2_y + \Delta^2_y, \qquad (6.3)$$

где: $\sum_{i=1}^{k} F^2_{y_i}$ – общая (нормированная) дисперсия – квадраты факторных нагрузок, соответствующих каждому исследуемому показателю i = от 1 до k;

Λ^2_y – специфическая (нормированная) дисперсия;

Δ^2_y – дисперсия (нормированная), обусловленная ошибкой измерения.

Отсюда следует, что квадраты факторных нагрузок показывают доли дисперсии измеряемого показателя, в том числе выходного параметра, приходящиеся на соответствующие факторы.

Если измеряемых показателей k, то можно построить матрицу факторных нагрузок $\mathbf{V}$, имеющую размерность $k * k$ и вычисляемую через корреляционную матрицу **R.**

Тогда ***фундаментальная теорема факторного анализа*** записывается в матричной форме так:

$$\mathbf{R} = \mathbf{V} * \mathbf{V}', \qquad (6.4)$$

где: **R** – матрица коэффициентов корреляции измеряемых показателей;

V – матрицу факторных нагрузок;

V′– транспонированная матрица.

Отсюда можно вычислить матрицу факторных нагрузок **V** из матрицы **R.**

6.2 Порядок выполнения

Исходную информацию для факторного анализа представим в виде выборки наблюдений объемом *m* * *k* над *k*-мерной случайной величиной, задаваемой матрицей наблюдений:

$$\mathbf{Y} = \begin{vmatrix} Y_{11} & Y_{12} & \cdot & \cdot & \cdot & Y_{1m} \\ Y_{21} & Y_{22} & \cdot & \cdot & \cdot & Y_{2m} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ Y_{k1} & Y_{k2} & \cdot & \cdot & \cdot & Y_{km} \end{vmatrix} \qquad (6.5)$$

Здесь *m* - количество результатов наблюдений над одним параметром, *k* - число контролируемых параметров, сюда входят все входные переменные и выходная переменная.

Рассчитаем все возможные коэффициенты корреляции r_{ij} между наблюдаемыми переменными для определения степени взаимосвязи между ними.

Из исходной матрицы **Y** (6.5) размерностью *m* * *k* получаем матрицу коэффициентов корреляции **R** путем умножения ее на соответствующую транспонированную матрицу и нормирования.

Новая матрица **R** (6.6), составленная из коэффициентов корреляции и имеющая размерность *k* * *k* , является основой факторного анализа:

$$\mathbf{R} = \begin{vmatrix} 1 & r_{12} & r_{13} & . & . & r_{1k} \\ r_{21} & 1 & r_{23} & . & . & r_{2k} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ r_{k1} & r_{k2} & r_{k3} & . & . & 1 \end{vmatrix} \qquad (6.6)$$

Величина и алгебраический знак коэффициента корреляции показывает, существует ли связь, и если существует, то каковы ее степень направление. Положительная связь, когда коэффициент корреляции имеет знак "плюс", свидетельствует, что чем выше одна переменная, тем выше и вторая. Отрицательная связь говорит о том, что чем выше одна переменная, тем меньше вторая. Нулевое или близкое к нулю значение показывает, что обе переменные изменяются независимо друг от друга.

Матрица факторных нагрузок **V** (6.7) рассчитывается из матрицы **R.**

В матрице **V** величина F_{ij} характеризует дисперсию, которую вносит *j*-й фактор в значение *i*-го показателя (например, выходной переменной)

$$\mathbf{V} = \begin{vmatrix} F_{11} & F_{12} & . & . & . & F_{1k} \\ F_{21} & F_{22} & . & . & . & F_{2k} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ F_{k1} & F_{k2} & . & . & . & F_{kk} \end{vmatrix} \qquad (6.7)$$

Столбец факторной матрицы **V** характеризует фактор и его влияние на все переменные. Строка характеризует переменную, ее наполненность различными факторами. В матрице факторных нагрузок строка, соответствующая выходной переменной, характеризует влияние на нее следующих факторов (слева направо) первого, второго, третьего. Для вычисления степени влияния факторов на выходную переменную необходимо соответствующие значения F_{ki} возвести в квадрат в строке, соответствующей выходной переменной.

Так как каждый фактор влияет на различные переменные, возникают трудности в содержательном толковании факторов. Поэтому в пакете ***STATISTICA for WINDOWS*** заложена возможность вращения матрицы факторных нагрузок несколькими способами для повышения доли отдельной переменной в каждом факторе.

6.3 Указания к выполнению

Для проведения факторного анализа выполним следующие действия:

- Составим таблицу исходных данных. Пусть в этой таблице входными переменными будут *Var1*, *Var2*, *Var3*, а выходной переменной, например, прибыль фирмы – переменная *Var4*. Проанализируем данные этой: таблицы оценим степень влияния переменных *Var1*, *Var2*, *Var3* на прибыль (переменная *Var4*).
- Запустим окно Factor Analysis посредством действий *Статистика/Многомерные исследующие методы/Анализ особенности* (рисунок 6.1) и выполним необходимые действия по вводу переменных. В качестве исследуемых переменных возьмем как входные *Var1*, *Var2*, *Var3*, так и выходную *Var4* переменные (рисунок 6.2).

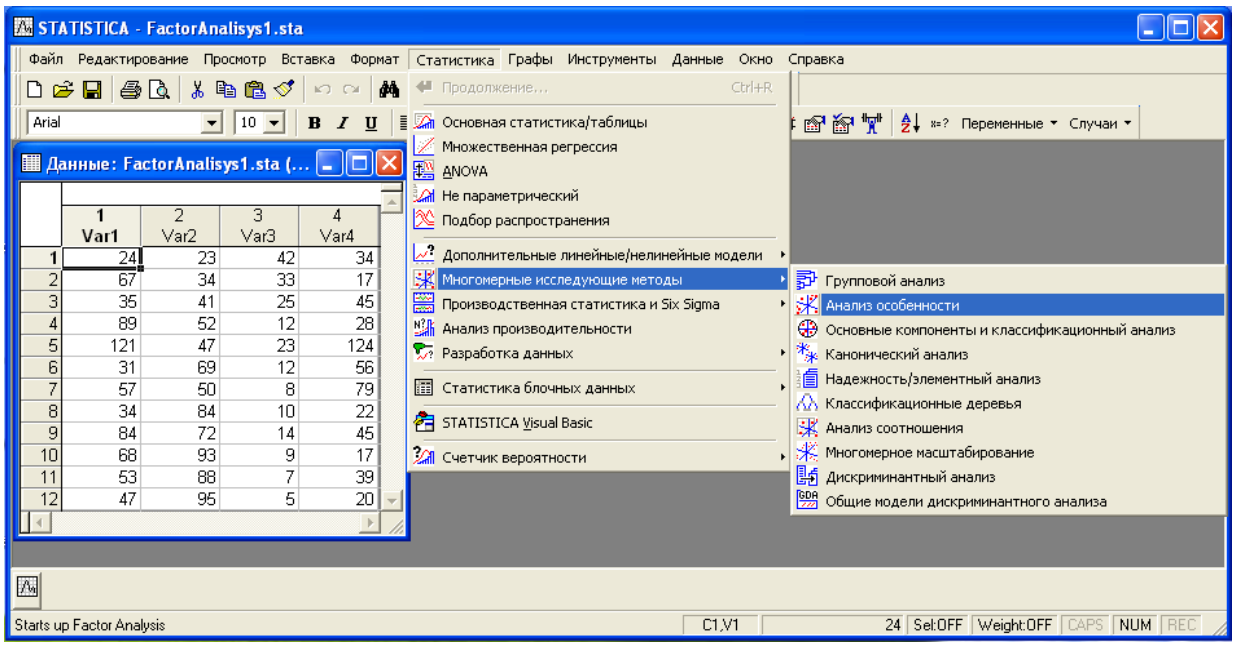


Рисунок 6.1. Окно Factor Analysis (Анализ особенности)

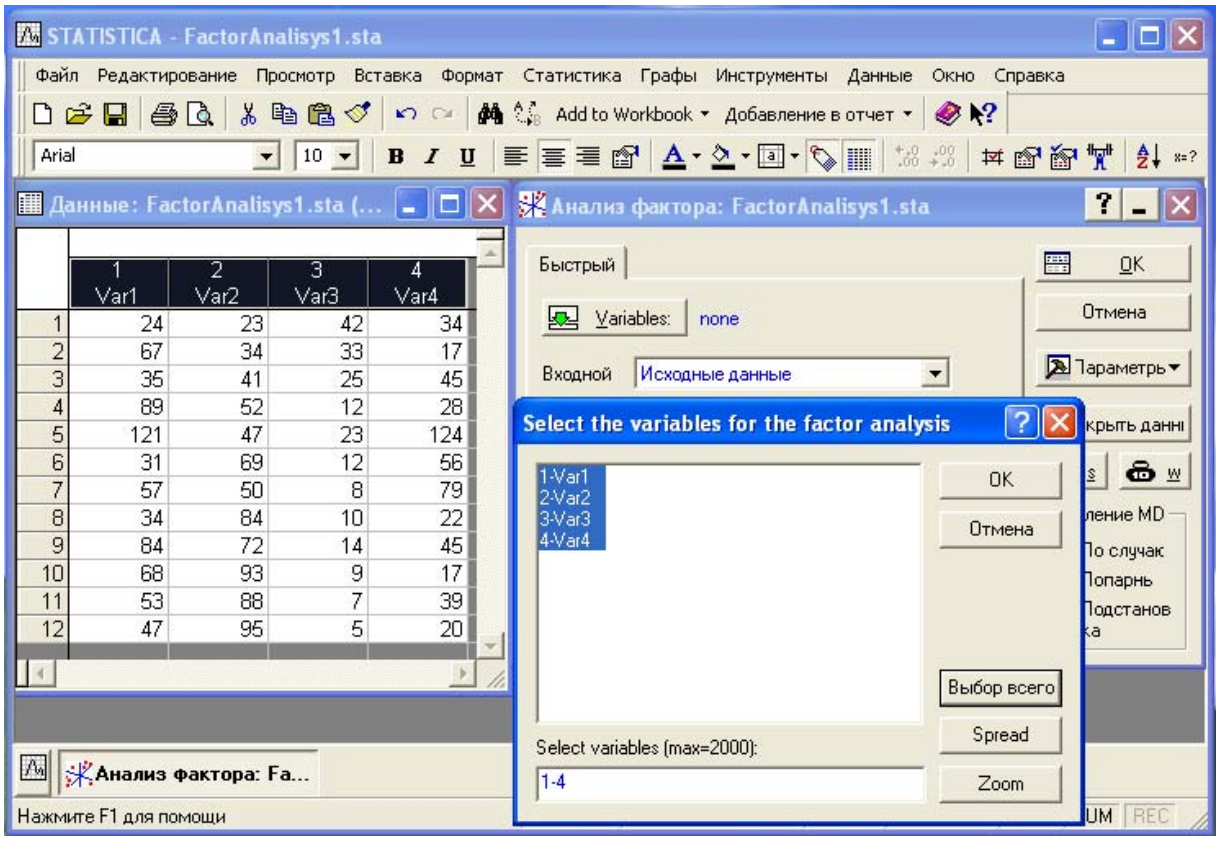


Рисунок 6.2 Окно Выбора переменных

- После запуска выполнения факторного анализа получим окно *Указать метод выборки фактора* (рисунок 6.3). Программа предоставляет возможность сразу выделить только наиболее значимые факторы, но, так как желательно видеть степени влияния на прибыль каждого фактора, оставим в результирующей таблице все факторы, для чего в окне *Указать метод выборки фактора/Расширенный* (рисунок 6.3) введем количество факторов (*Максимальное*) - 4, а минимальный порог (*Миним.*) установим достаточно низким, например, 0.05. В качестве метода выборки факторов установим метод *Основные компоненты*.

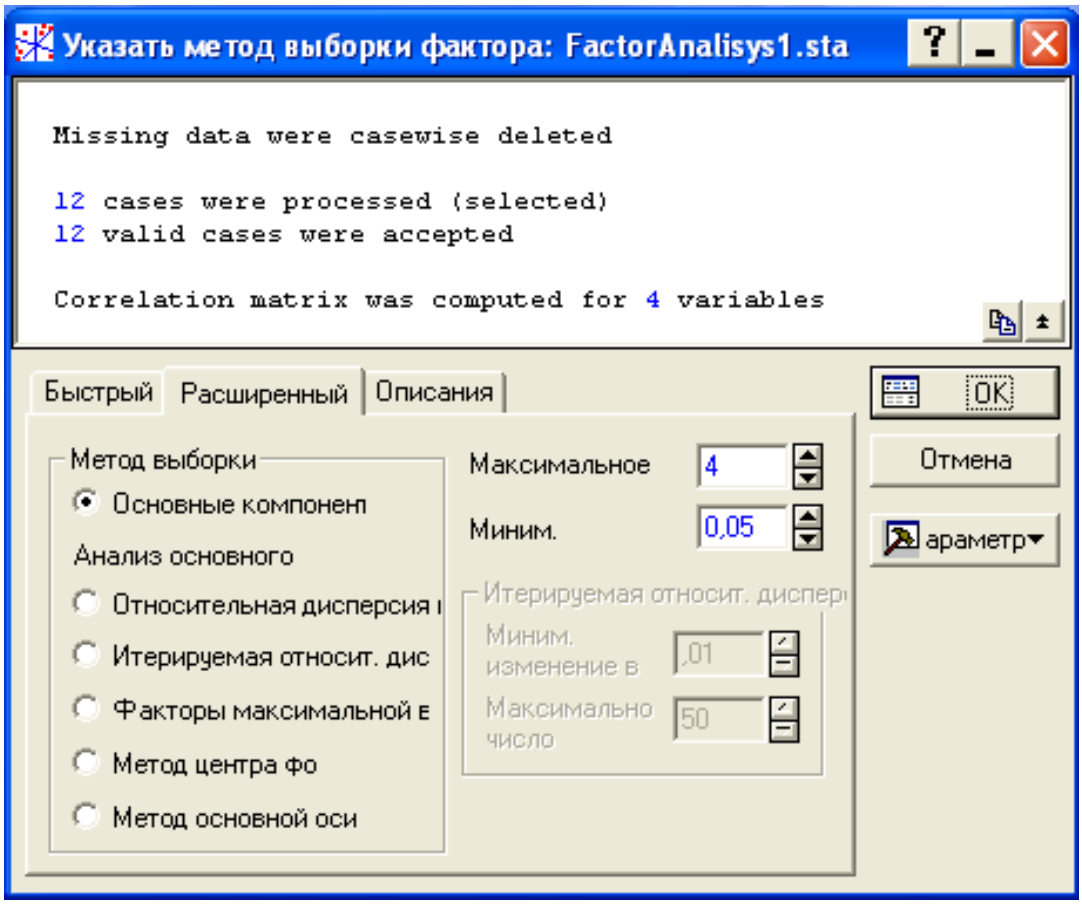


Рисунок 6.3 Определение метода выборки факторов

- После запуска выполнения выделения факторов появится окно *Результаты анализа факторов* (рисунок 6.4).

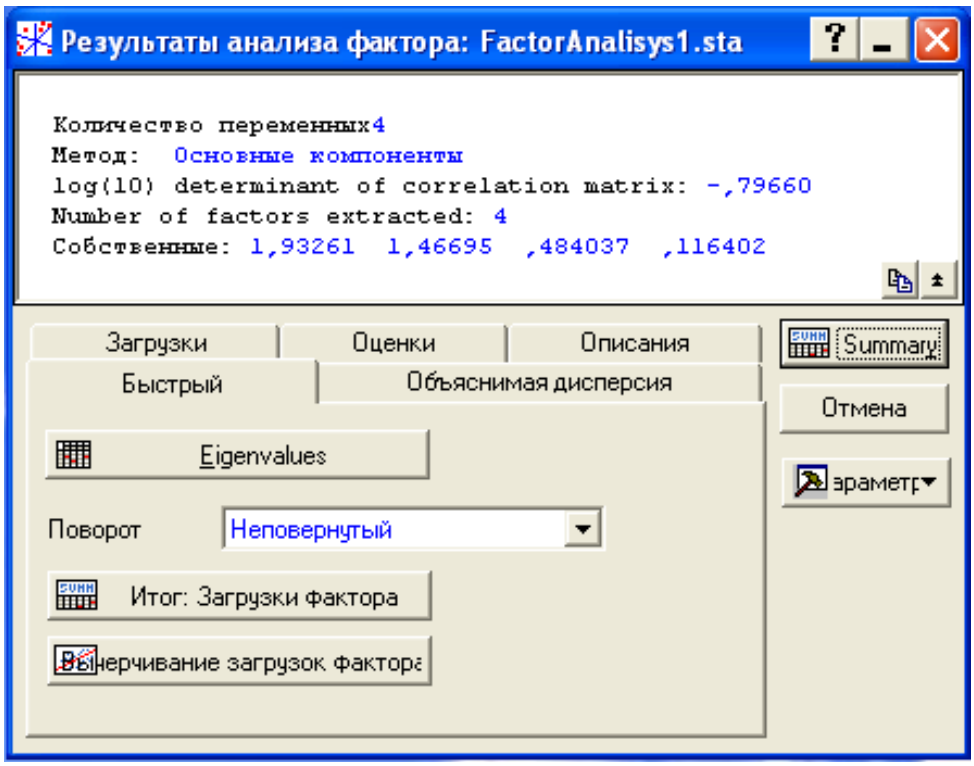


Рисунок 6.4 Окно результатов факторного анализа

- Так как нас интересует степени влияния переменных *Var1*, *Var2*, *Var3* на переменную *Var4*, запустим *Итог: Загрузки фактора*. В результате получим *матрицу факторных нагрузок* (рисунок 6.5).

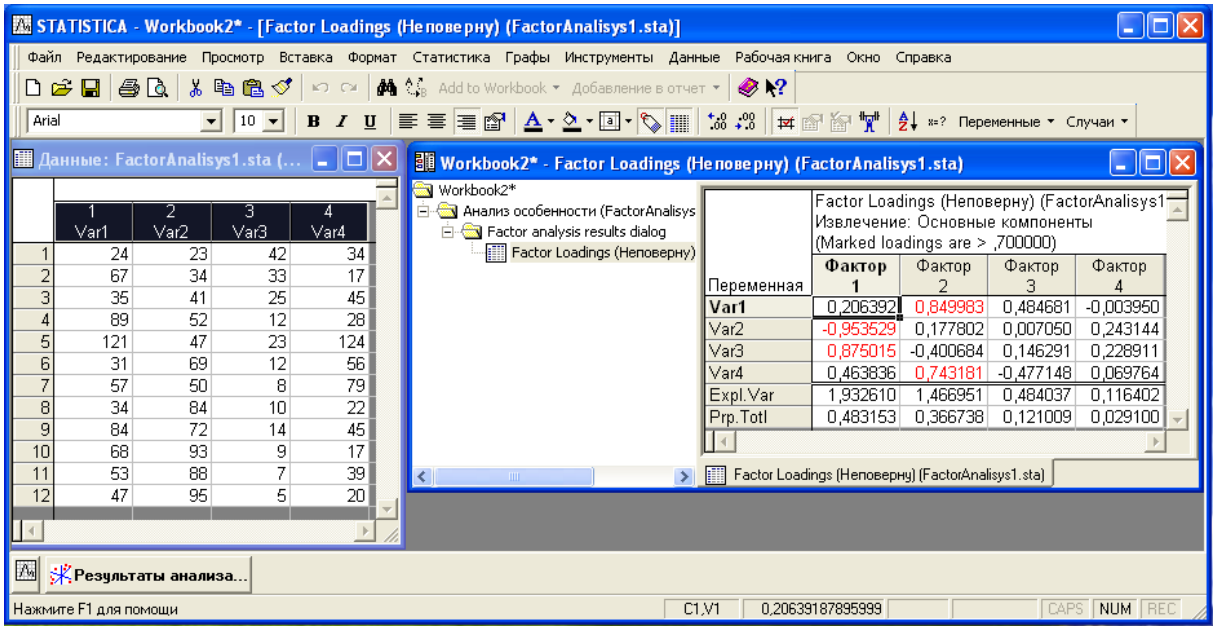


Рисунок 6.5 Матрица факторных нагрузок

- Влияние на прибыль (*Var4*) разных факторов вычислим путем возведения в квадрат значений, находящихся в строке *Var4*. Сумма полученных долей влияния факторов на переменную, рассчитываемых как квадраты факторных нагрузок, равна 1.
- В рассматриваемом примере наибольшее влияние на переменную *Var4* оказывает *Фактор 2*. Теперь рассмотрим наполненность *Фактора 2* переменными (кроме переменной *Var4*, так как эта переменная отражает внутреннюю дисперсию самой переменной *Var4)*. Наибольшую наполненность *Фактора 2* создают переменные *Var1*, затем *Var3*. Но Влияние *Фактора 2* на переменную *Var4* не очень велико (с учетом возведения в квадрат значений в строке 4), поэтому попробуем повернуть матрицу факторных нагрузок. Для этого в окне *Результаты анализа фактора* (рисунок 6.6) установим один из способов поворота матрицы, например, *Варимаксный нормализованный*.

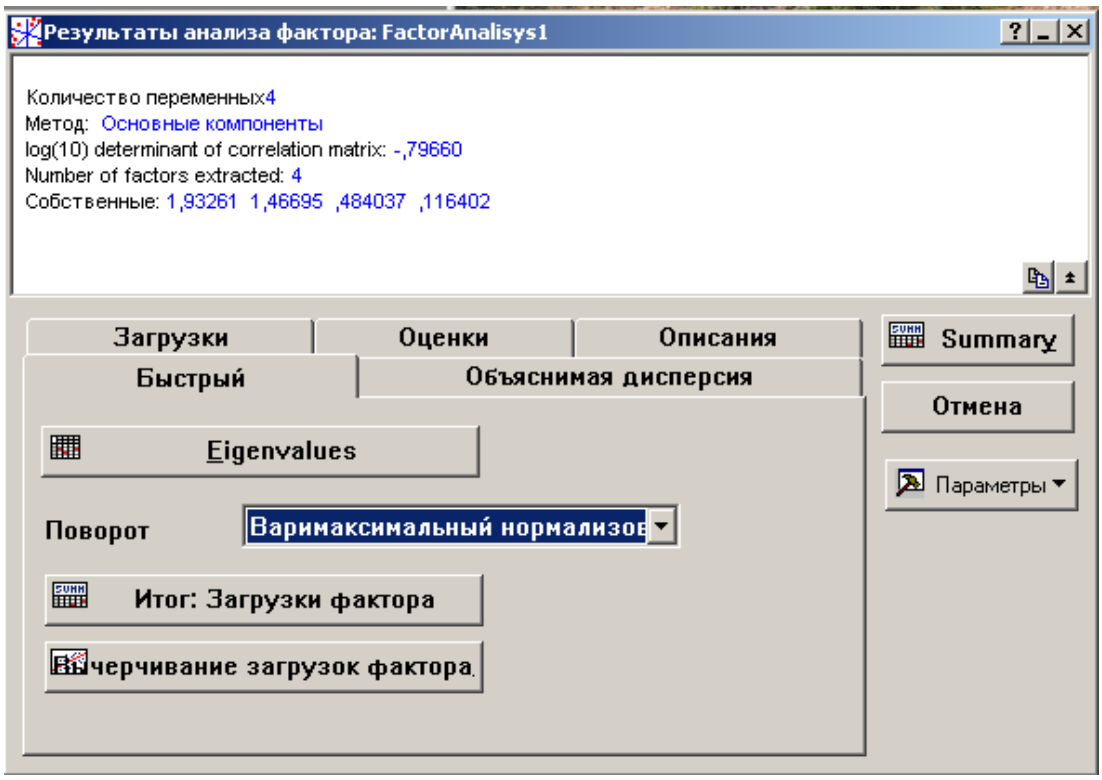


Рисунок 6.6 Задание способа поворота матрицы факторных нагрузок

- В результате получим другую матрицу факторных нагрузок (рисунок 6.7). В этой матрице влияние одного из факторов *Фактор 3* на переменную *Var4* является доминирующим. Поэтому результаты в этой матрице более достоверны, нежели в предыдущей. Наибольшую наполненность *Фактора 3* создают переменные *Var1*, затем *Var2*.

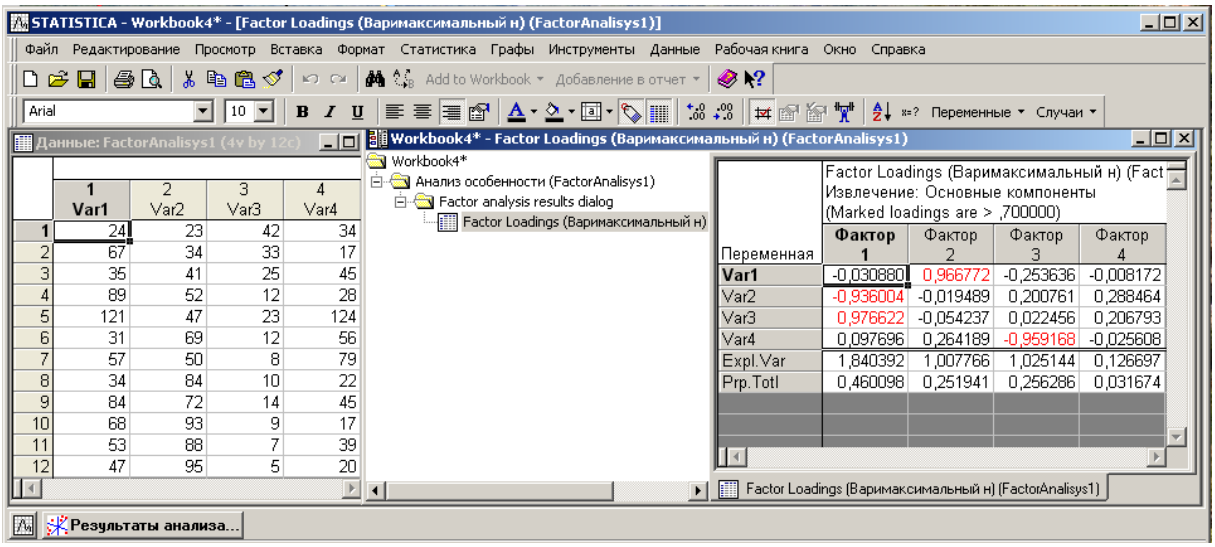


Рисунок 6.7 Повернутая матрица факторных нагрузок

7 Каким фирмам можно доверять?

7.1 Основы кластерного анализа

Кластерный анализ предназначен для разбиения множества объектов на заданное или неизвестное число классов на основании некоторого критерия классификации (***cluster*** – гроздь, группа элементов, характеризуемых каким-то общим свойством). Поэтому, название раздела: «Каким фирмам можно доверять?» – это только частный случай кластерного анализа.

В кластерном анализе множество данных нужно разбить на некоторое число групп так, чтобы, с одной стороны, каждый объект принадлежал только к одной группе и, с другой, – объекты, входящие в одну группу, были максимально схожими, а разные группы – были разнородными. Т.е. каждая совокупность объектов должна группироваться вокруг некоего центра группы (рисунок 7.1).

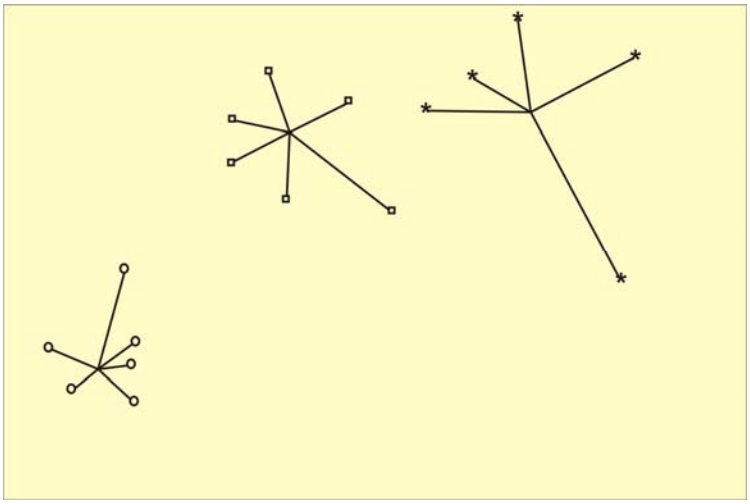


Рисунок 7.1 Группировка объектов вокруг центров групп

Для пояснения сущности кластерного анализа рассмотрим примеры.

Пример 1.

Вы начальник кредитного отдела банка. Столкнувшись с невозвратами кредитов, Вы решаете впредь выдавать кредиты лишь фирмам, которые «схожи» с теми, которые хорошо себя зарекомендовали, и не выдавать тем, которые «схожи» с неплательщиками. Для классификации фирм можно собрать показатели их деятельности и провести кластерный анализ. В реальности работа фирм характеризуется множеством показателей, но в данном примере мы упростим задачу – оставим только два показателя: затраты *x* и прибыль *y* за предыдущий период (Таблица 7.1).

Таблица 7.1 Показатели работы фирм

	1 Номер фирмы	2 Затраты *X*	3 Прибыль *Y*
1	1	4	2
2	2	6	10
3	3	5	7
4	4	12	3
5	5	17	4
6	6	3	10
7	7	6	1
8	8	6	3
9	9	15	1
10	10	15	4
11	11	5	4
12	12	3	8
13	13	13	5
14	14	15	3
15	15	5	9
16	16	8	3
17	17	6	4
18	18	14	3
19	19	4	8
20	20	4	3

Построим двумерный график распределения показателей работы фирм в координатах *x* и *y* (рисунок 7.2).

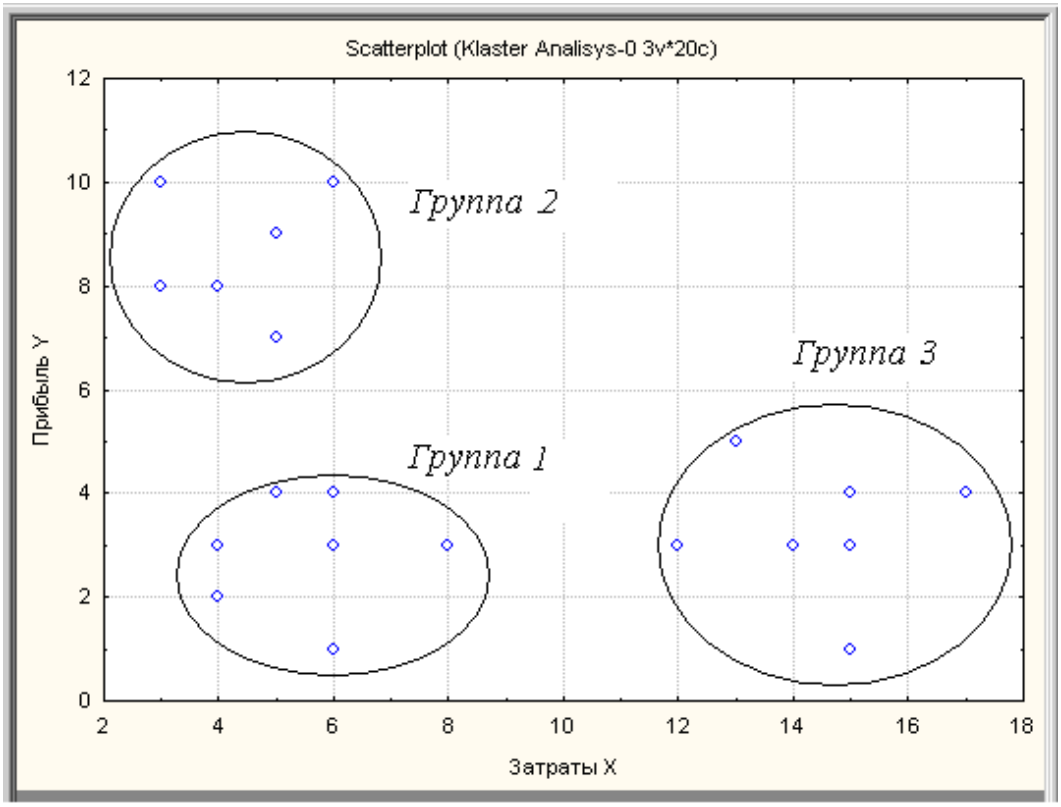


Рисунок 7.2 Распределение показателей работы фирм

Из графика (рисунрк 7.2) видно, что все фирмы можно разделить на три части, которые образуют достаточно тесные группы. Группа 1 (1, 20, 7, 16, 8, 17, 11) характеризуется небольшими затратами, но и прибыль у нее невысока. Группа 2 (2, 15, 6, 3, 12, 19) характеризуется небольшими затратами и высокой прибылью. У группы 3 (4, 13, 5, 9, 10, 14, 18) большие затраты и невысокая прибыль. Таким образом, можно сделать вывод: группу 2 поддержать, группе один кредиты выдавать только после тщательного обоснования, а группе 3 кредиты не выдавать. Распределение фирм на группы с помощью кластерного анализа (о процедуре его выполнения ниже) показано на рисунке 7.3.

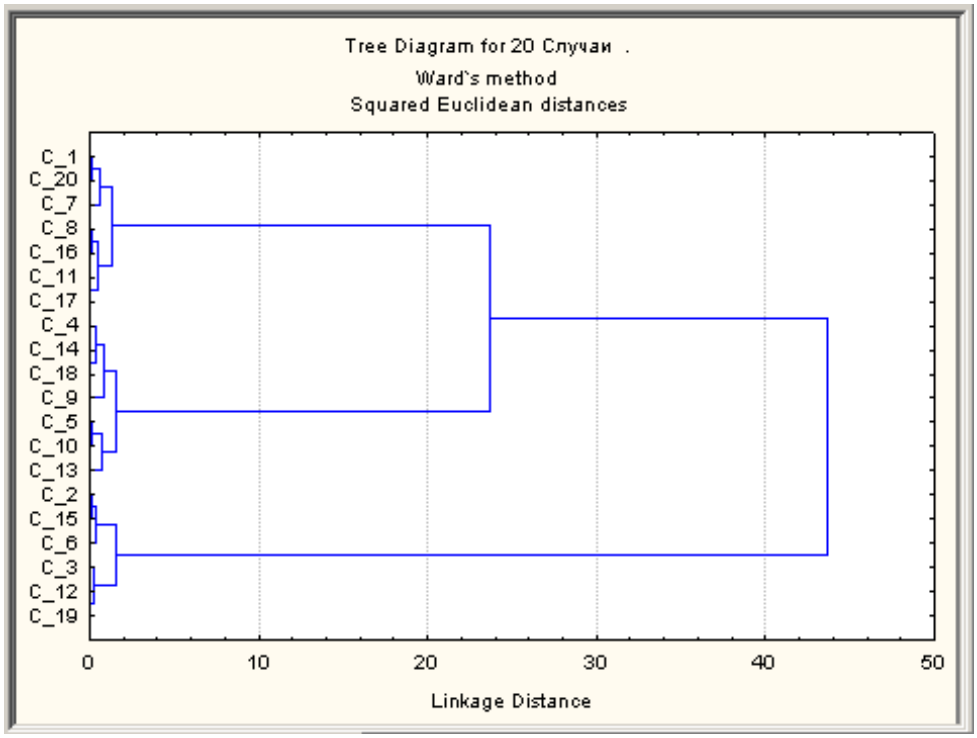


Рисунок 7.3 Разделение фирм на три группы (кластеры)

Показатели работы фирм с рекомендациями по целесообразности предоставления кредитов представлена в таблице 7.2.

Таблица 7.2. Показатели работы фирм с рекомендациями.

	1 Номер фирмы	2 Затраты *X*	3 Прибыль *Y*	4 Рекомендация
1	1	4	2	Присмотреться
2	2	6	10	Поддержать
3	3	5	7	Поддержать
4	4	12	3	Не выдавать
5	5	17	4	Не выдавать
6	6	3	10	Поддержать
7	7	6	1	Присмотреться
8	8	6	3	Присмотреться
9	9	15	1	Не выдавать
10	10	15	4	Не выдавать
11	11	5	4	Присмотреться
12	12	3	8	Поддержать
13	13	13	5	Не выдавать
14	14	15	3	Не выдавать
15	15	5	9	Не выдавать
16	16	8	3	Присмотреться
17	17	6	4	Присмотреться
18	18	14	3	Не выдавать
19	19	4	8	Поддержать
20	20	4	3	Присмотреться

Особенностью данного примера является то, что фактически изначально было известно, на сколько групп нужно делить все фирмы: хорошие, плохие и середнячки, т.е. на 3 группы. Все данные о фирмах были двумерными, т.е. затраты и прибыль, причем эти данные имели одинаковую размерность.

Пример 2.

Пусть имеется набор цифр, т.е. однородный случай: 8, 4, 2, 2, 4, 8, 2, 6, 4, 8, 2. В отличие от предыдущего случая заранее не известно, на сколько групп следует разделить весь набор. Примем в качестве критерия классификации сумму квадратов отклонений $W = \sum (x_i - \bar{x})^2 \to \min$. Для исходного набора данных W = 64,27.

Теперь нужно разбить всю совокупность цифр на группы таким образом, чтобы W = min. В данном простейшем случае очевидно, что лучшим решением будет 4 группы цифр:

A1 = {8, 8, 8};

A2 = {4, 4, 4};

A3 = {2, 2, 2, 2};

A4 = {6}.

В таком случае W1 = W2 = W3 = W4= 0. Все легко и просто?

Но в реальности экономические объекты характеризуются множеством показателей, имеющих различные размерность и численные значения. Не всегда известно, на сколько кластеров целесообразно разбивать совокупность исследуемых объектов. Часть объектов можно с примерно одинаковой обоснованностью отнести к тому или иному кластеру. Поэтому существуют несколько методов обработки данных.

7.2 Порядок выполнения

- По экспериментальным данным строится матрица исходных данных: $X[n,k]$, где *n – количество объектов, k – количество показателей.*
- Исходная матрица стандартизуется (данные переводятся в безразмерную относительную форму, чтобы можно было соизмерять разнородные показатели). Стандартизованное значение x_i^* рассчитывается по формуле: $x_i^* = \frac{x_i - \bar{x}}{S_h}$, (7.1)

  где x_i - исходные данные; $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ - выборочное среднее;

  $S_h = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ - среднее квадратическое отклонение.
- На 1 этапе задается начальное разбиение множества объектов на классы и определяется математический критерий классификации.
- Задаются метрики расстояния:

- евклидово расстояние: $\rho(x_1, x_z) = \sqrt{\sum_{m=1}^{k}(x_{im} - x_{lm})^2}$, (7.2)

  где *i, z* = 1,2,3…*n;*

- или квадратическое евклидово расстояние:

$$\rho(x_1, x_z) = \sum_{m=1}^{k}(x_{im} - x_{lm})^2. \quad (7.3)$$

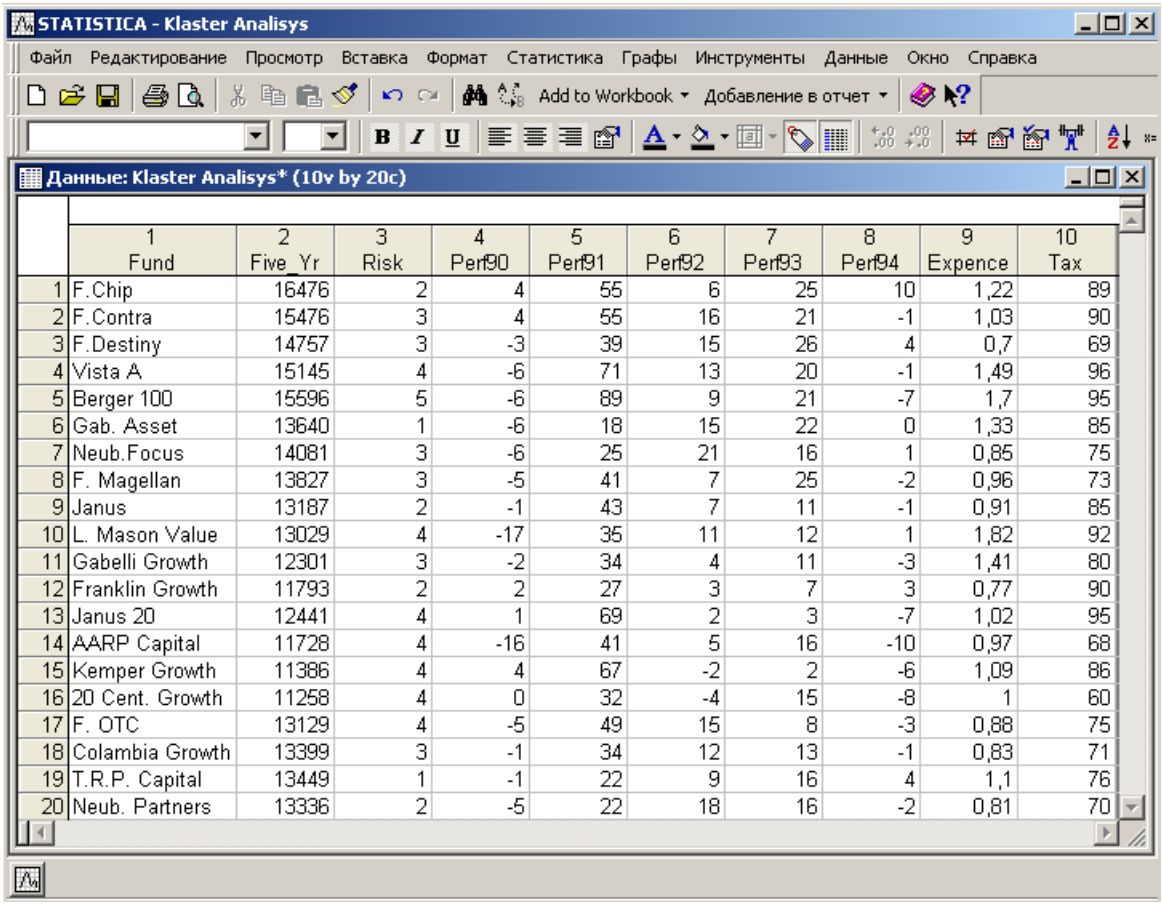
- Целевая функция: внутригрупповая сумма квадратов должна минимально увеличиваться при объединении кластеров.

- На 2 этапе объекты переносятся из класса в класс до достижения экстремума критерия классификации.

7.3 Указания к выполнению

В качестве примера проведения кластерного анализа рассмотрим данные о результатах работы 20 инвестиционных фондов [3]. Разобьем эти фонды на три группы: те, которым можно доверять, те, которым доверять не следует и середнячков. Для проведения кластерного анализа выполним следующие действия:

- В пакете ***STATISTICA for WINDOWS*** создадим базу данных (рисунок 7.4).



	1 Fund	2 Five_Yr	3 Risk	4 Perf90	5 Perf91	6 Perf92	7 Perf93	8 Perf94	9 Expence	10 Tax
1	F.Chip	16476	2	4	55	6	25	10	1,22	89
2	F.Contra	15476	3	4	55	16	21	-1	1,03	90
3	F.Destiny	14757	3	-3	39	15	26	4	0,7	69
4	Vista A	15145	4	-6	71	13	20	-1	1,49	96
5	Berger 100	15596	5	-6	89	9	21	-7	1,7	95
6	Gab. Asset	13640	1	-6	18	15	22	0	1,33	85
7	Neub.Focus	14081	3	-6	25	21	16	1	0,85	75
8	F. Magellan	13827	3	-5	41	7	25	-2	0,96	73
9	Janus	13187	2	-1	43	7	11	-1	0,91	85
10	L. Mason Value	13029	4	-17	35	11	12	1	1,82	92
11	Gabelli Growth	12301	3	-2	34	4	11	-3	1,41	80
12	Franklin Growth	11793	2	2	27	3	7	3	0,77	90
13	Janus 20	12441	4	1	69	2	3	-7	1,02	95
14	AARP Capital	11728	4	-16	41	5	16	-10	0,97	68
15	Kemper Growth	11386	4	4	67	-2	2	-6	1,09	86
16	20 Cent. Growth	11258	4	0	32	-4	15	-8	1	60
17	F. OTC	13129	4	-5	49	15	8	-3	0,88	75
18	Colambia Growth	13399	3	-1	34	12	13	-1	0,83	71
19	T.R.P. Capital	13449	1	-1	22	9	16	4	1,1	76
20	Neub. Partners	13336	2	-5	22	18	16	-2	0,81	70

Рисунок 7.4 База данных инвестиционных фондов

- Кластерный анализ позволяет провести многомерную группировку объектов. Но еще до выполнения кластерного анализа можно визуально рассмотреть двумерное или трехмерное распределение

объектов. Наиболее важными показателями являются прибыль за 5 лет (*Five Yr*) и затраты (*Expence*). Именно соотношение *Прибыль/Затраты* является важнейшим показателем. Необходимо заметить, что двумерное распределение может показать только приблизительную картину, тем не менее, построим график двумерного распределения, для чего выполним действия *Графы/Точечные вычерчивания/Расширенный.* В результате получим окно (рисунок 7.5).

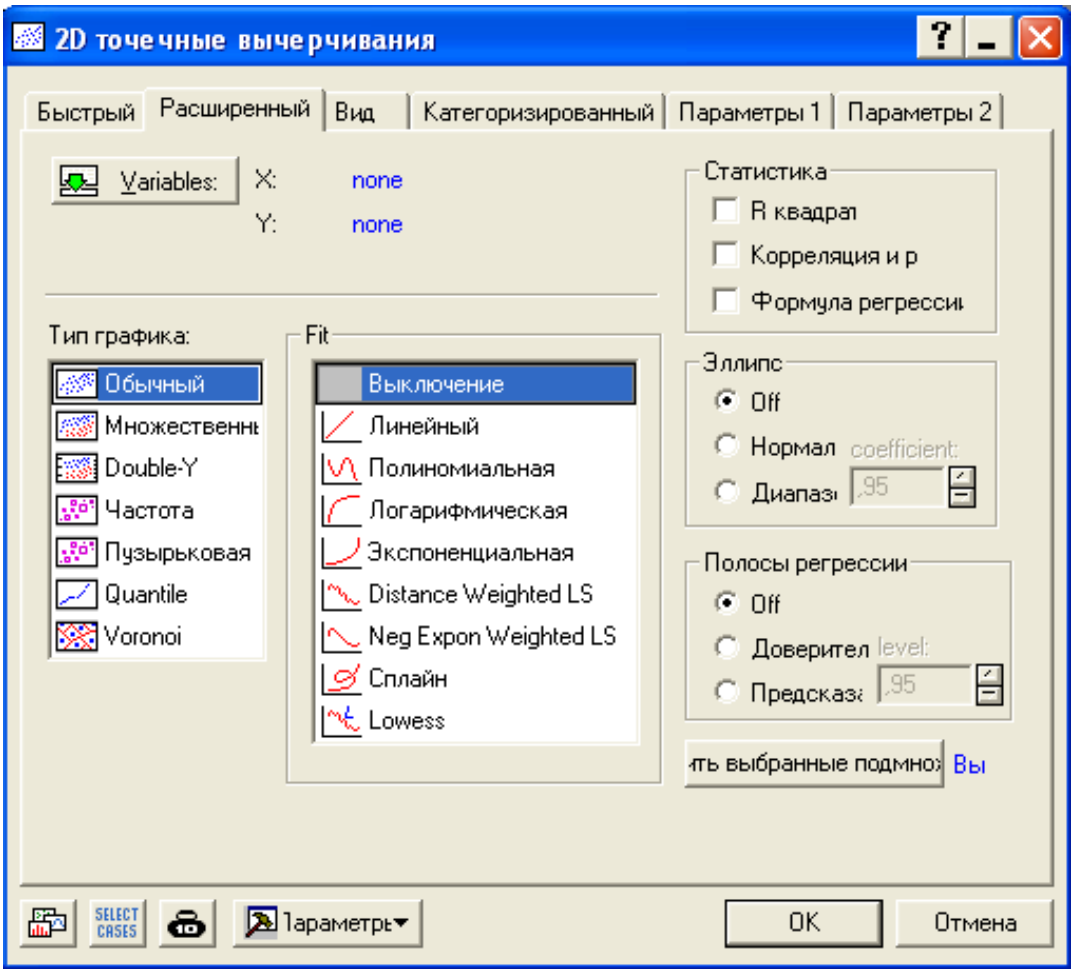


Рисунок 7.5 Окно *2 D точечные вычерчивания/Расширенный*

- В окне *2 D точечные вычерчивания* выберем *тип графика: Обычный* и вид зависимости (*Fit*): *Выключение* (рисунок 7.5). Установим переменные *Variables*: *x- Expence, y- Five Yr*.

- Для обозначения точек на будущем графике прейдем в окне *2 D точечные вычерчивания* (рисунок 7.5) с раздела *Расширенный* в раздел *Параметры 1*, получим окно (рисунок 7.6)

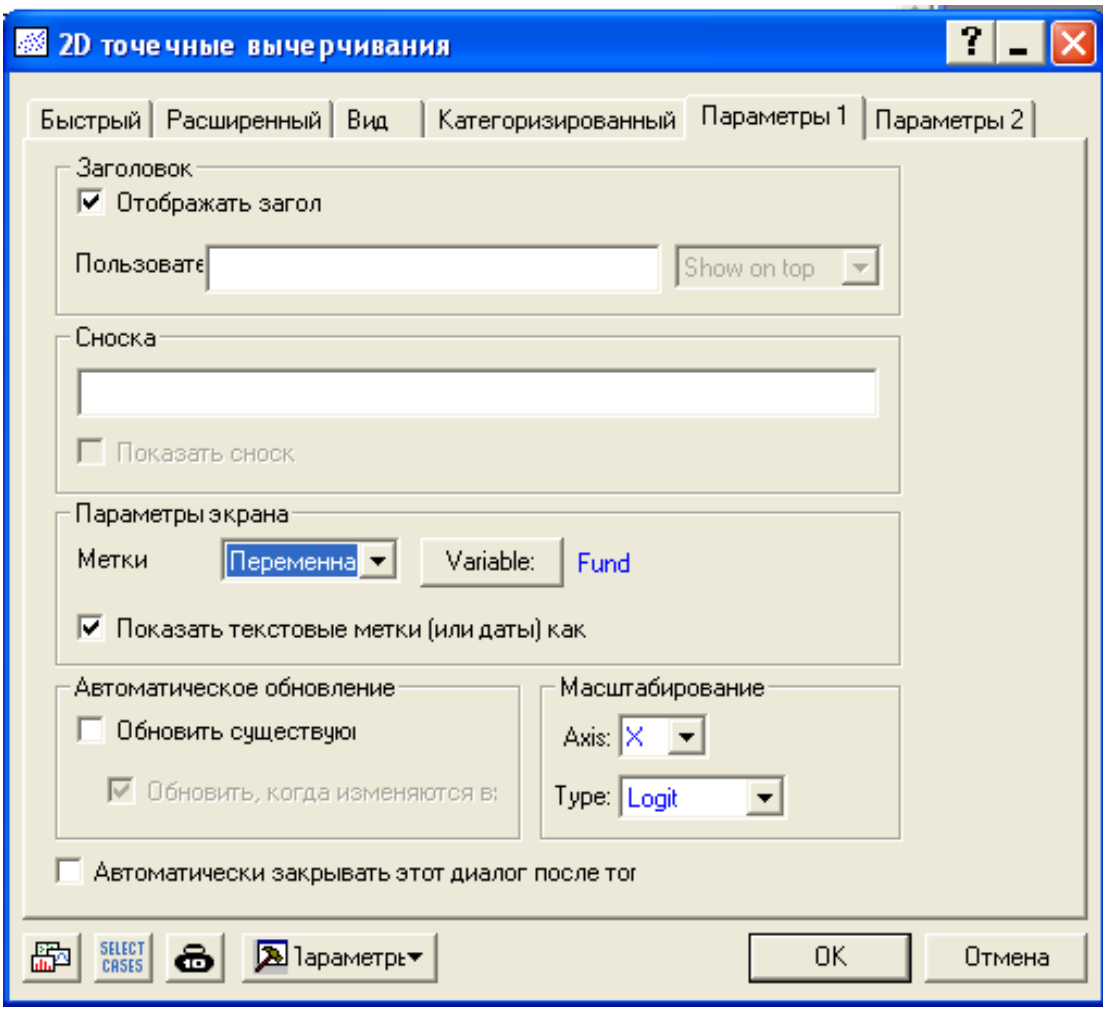


Рисунок 7.6 Окно *2 D точечные вычерчивания/Параметры 1*

- В окне (рисунок 7.6) можно задать *метки*: *случаи* или *переменные*. Установим *Переменные,* тогда в ячейке *Variable* название фондов: *Fund*. Подтвердив заданные параметры, получим график Scatterplot с распределением объектов в пространстве *x- Expence, y- Five Yr* (рисунок 7.7).

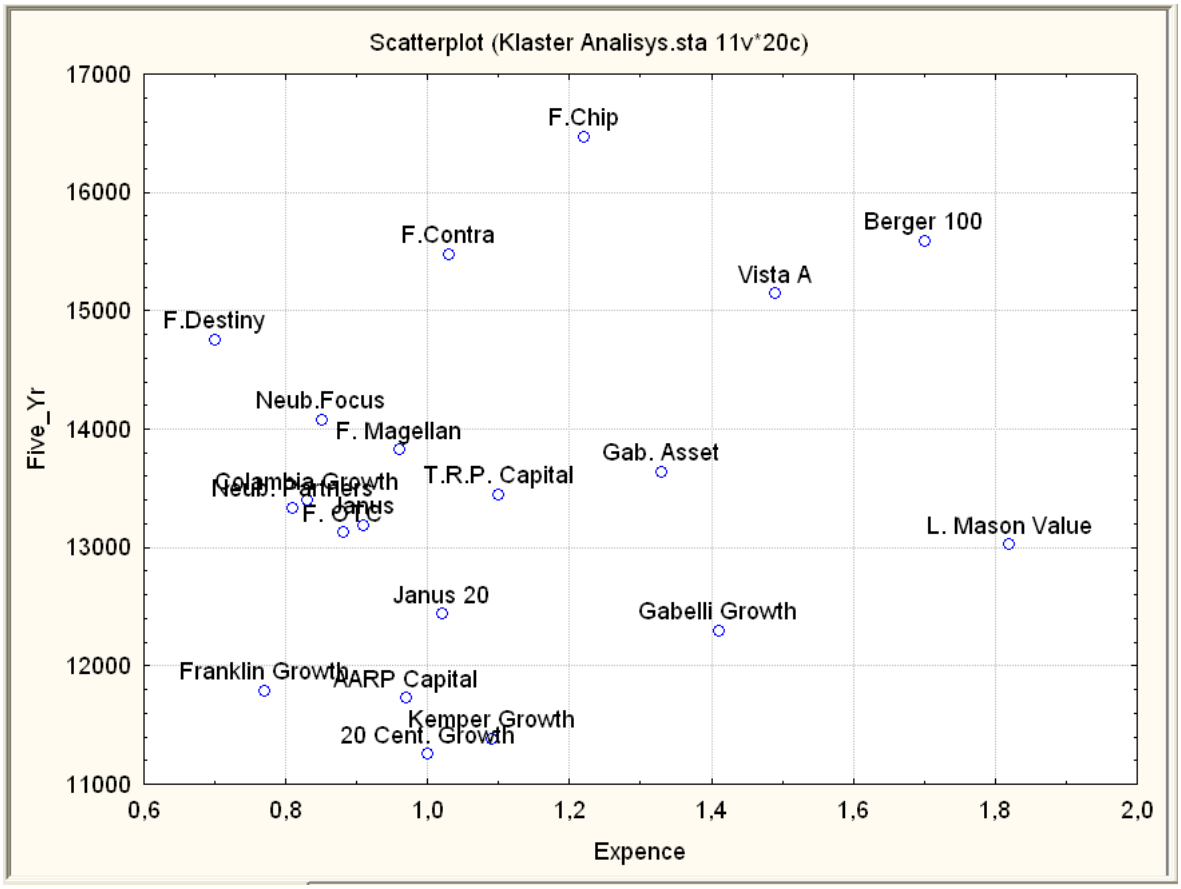


Рисунок 7.7 Распределением объектов в пространстве *x- Expence, y- Five Yr*

- Для выполнения кластерного анализа нужно выполнить предварительную обработку исходных данных в таблице (рисунок7.4), так как они имеют различную размерность и существенно различаются по модулю численных значений. Поэтому необходимо их перевести в относительные величины, для чего каждое значение разделить на среднее по колонке (данные одной размерности находятся в колонках). Т.е. нужно провести стандартизацию исходных данных. Для этого выделим всю часть таблицы с численными значениями и выполним следующую последовательность действий:

  Редактирование/Заполнение /стандартизация блока/Стандартизация столбцов (рисунок 7.8).

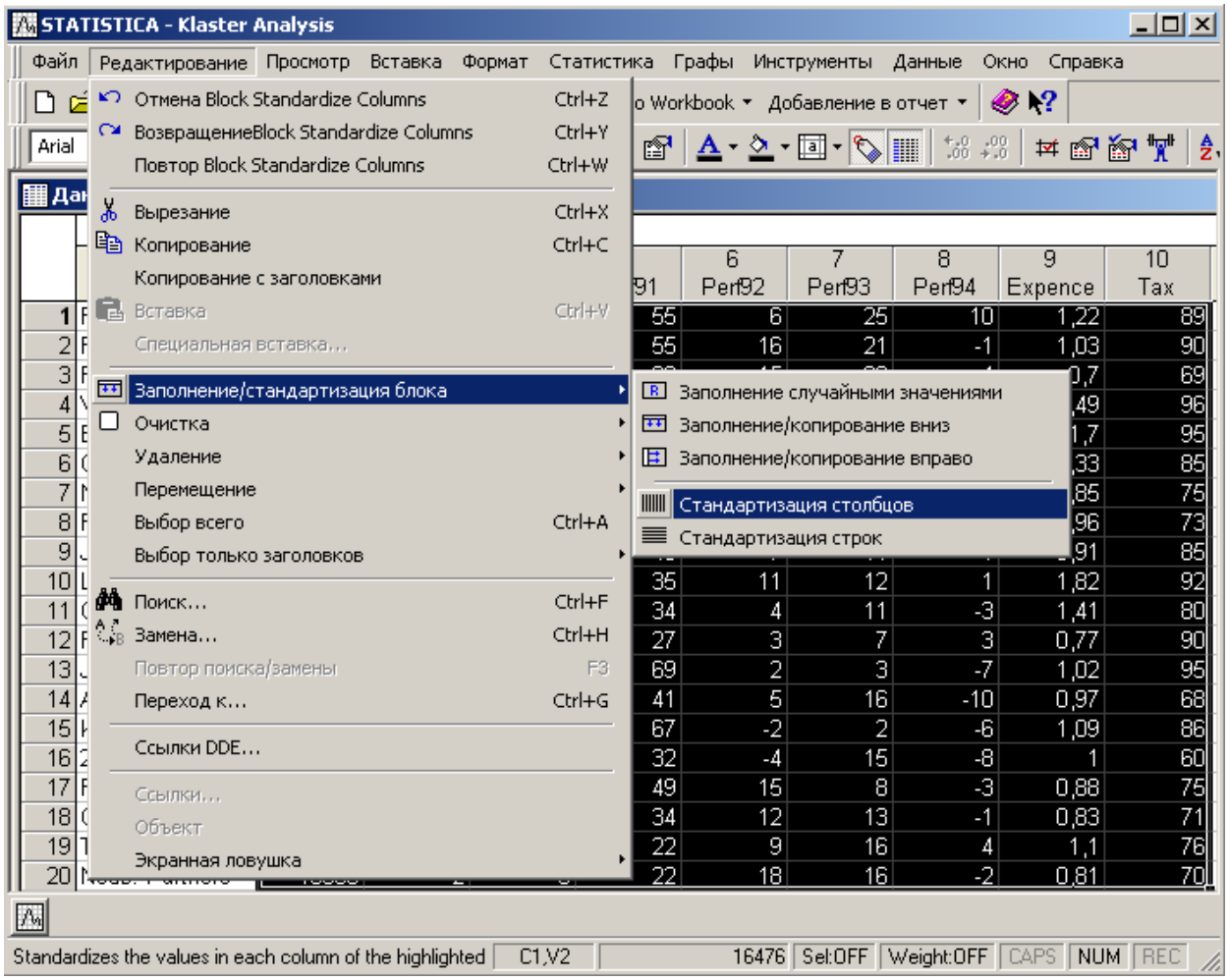
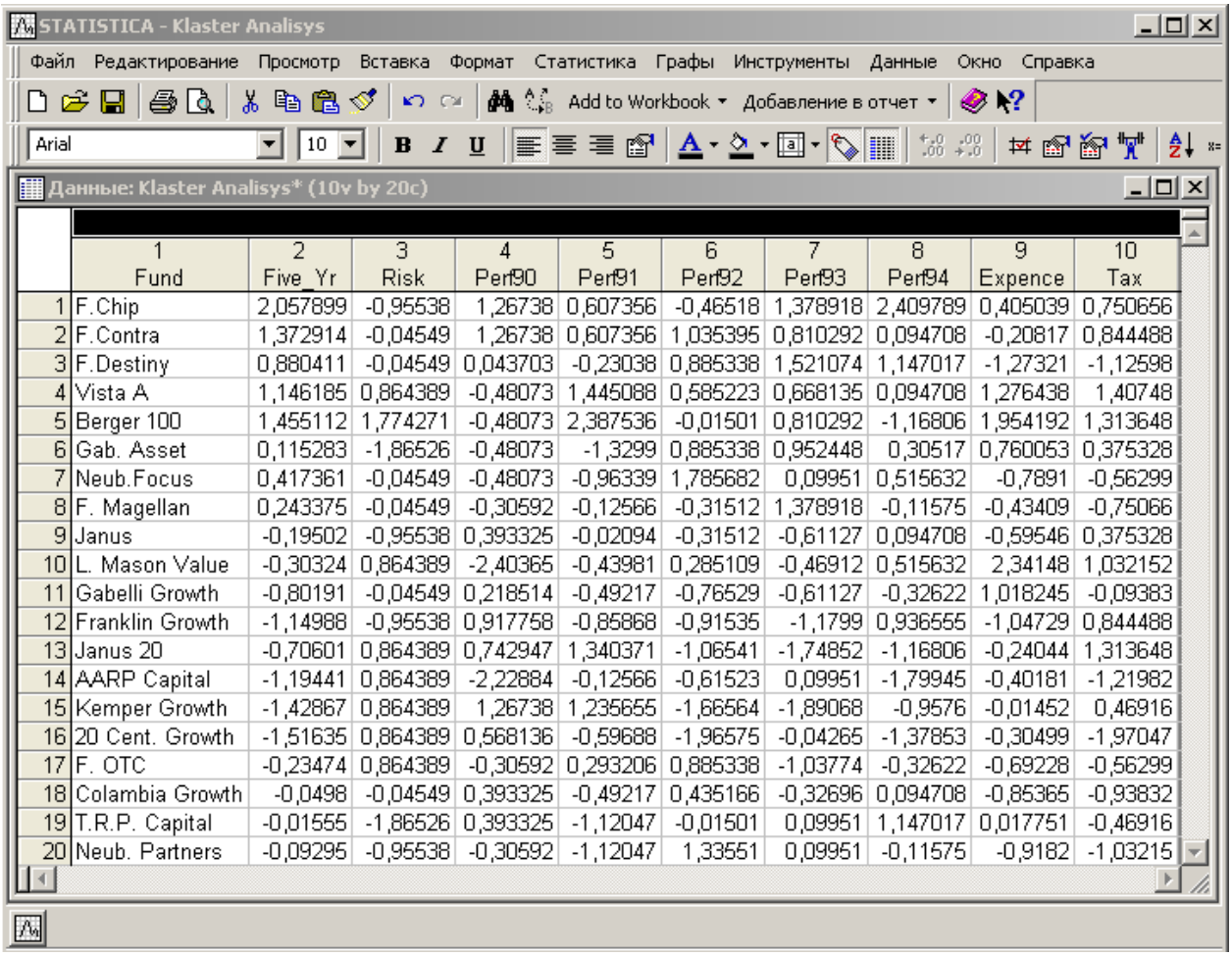


Рисунок 7.8 Последовательность действий стандартизации

- В результате получим стандартизованную базу данных (рисунок 7.9).



	1 Fund	2 Five_Yr	3 Risk	4 Perf90	5 Perf91	6 Perf92	7 Perf93	8 Perf94	9 Expence	10 Tax
1	F.Chip	2,057899	-0,95538	1,26738	0,607356	-0,46518	1,378918	2,409789	0,405039	0,750656
2	F.Contra	1,372914	-0,04549	1,26738	0,607356	1,035395	0,810292	0,094708	-0,20817	0,844488
3	F.Destiny	0,880411	-0,04549	0,043703	-0,23038	0,885338	1,521074	1,147017	-1,27321	-1,12598
4	Vista A	1,146185	0,864389	-0,48073	1,445088	0,585223	0,668135	0,094708	1,276438	1,40748
5	Berger 100	1,455112	1,774271	-0,48073	2,387536	-0,01501	0,810292	-1,16806	1,954192	1,313648
6	Gab. Asset	0,115283	-1,86526	-0,48073	-1,3299	0,885338	0,952448	0,30517	0,760053	0,375328
7	Neub.Focus	0,417361	-0,04549	-0,48073	-0,96339	1,785682	0,09951	0,515632	-0,7891	-0,56299
8	F. Magellan	0,243375	-0,04549	-0,30592	-0,12566	-0,31512	1,378918	-0,11575	-0,43409	-0,75066
9	Janus	-0,19502	-0,95538	0,393325	-0,02094	-0,31512	-0,61127	0,094708	-0,59546	0,375328
10	L. Mason Value	-0,30324	0,864389	-2,40365	-0,43981	0,285109	-0,46912	0,515632	2,34148	1,032152
11	Gabelli Growth	-0,80191	-0,04549	0,218514	-0,49217	-0,76529	-0,61127	-0,32622	1,018245	-0,09383
12	Franklin Growth	-1,14988	-0,95538	0,917758	-0,85868	-0,91535	-1,1799	0,936555	-1,04729	0,844488
13	Janus 20	-0,70601	0,864389	0,742947	1,340371	-1,06541	-1,74852	-1,16806	-0,24044	1,313648
14	AARP Capital	-1,19441	0,864389	-2,22884	-0,12566	-0,61523	0,09951	-1,79945	-0,40181	-1,21982
15	Kemper Growth	-1,42867	0,864389	1,26738	1,235655	-1,66564	-1,89068	-0,9576	-0,01452	0,46916
16	20 Cent. Growth	-1,51635	0,864389	0,568136	-0,59688	-1,96575	-0,04265	-1,37853	-0,30499	-1,97047
17	F. OTC	-0,23474	0,864389	-0,30592	0,293206	0,885338	-1,03774	-0,32622	-0,69228	-0,56299
18	Colambia Growth	-0,0498	-0,04549	0,393325	-0,49217	0,435166	-0,32696	0,094708	-0,85365	-0,93832
19	T.R.P. Capital	-0,01555	-1,86526	0,393325	-1,12047	-0,01501	0,09951	1,147017	0,017751	-0,46916
20	Neub. Partners	-0,09295	-0,95538	-0,30592	-1,12047	1,33551	0,09951	-0,11575	-0,9182	-1,03215

Рисунок 7.9 Стандартизованная база данных инвестиционных фондов

- Для запуска кластерного анализа выполним следующую последовательность действий *Статистика/Многомерные исследующие методы/Групповой анализ* (рисунок 7.10).

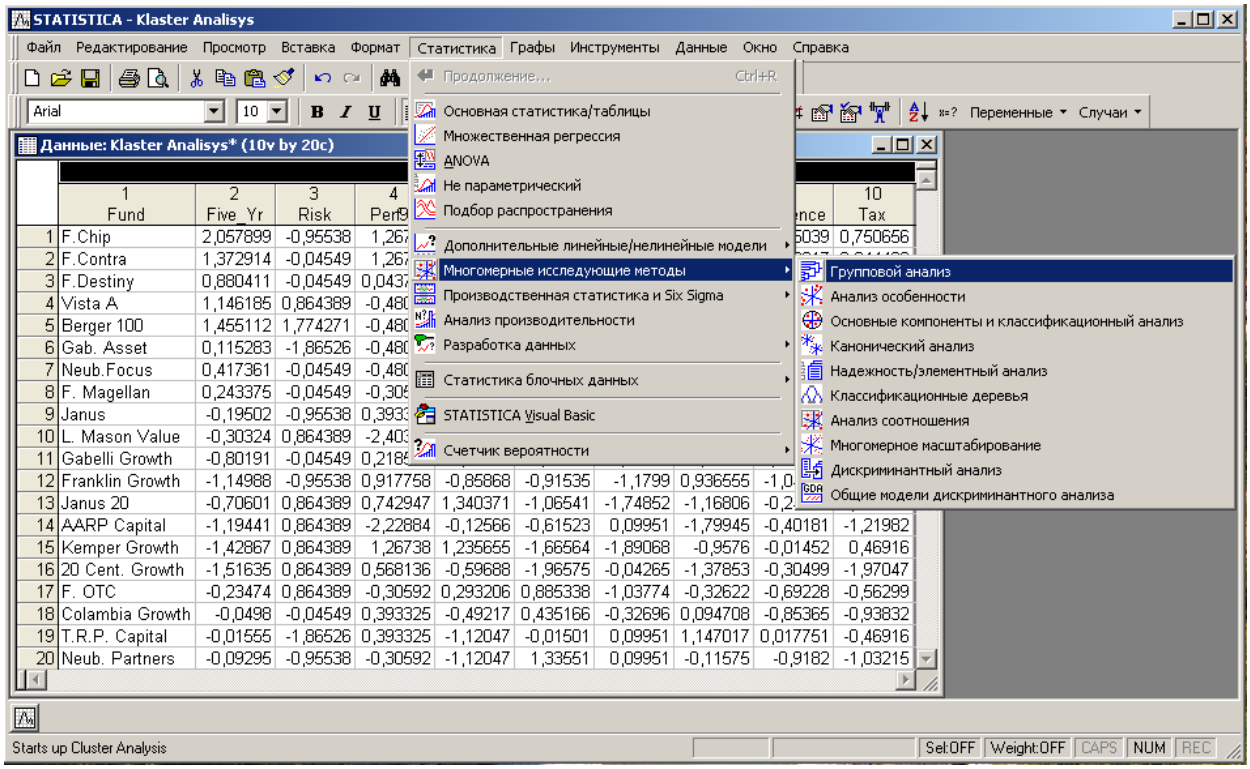


Рисунок 7.10 Начало кластерного анализа

- Запустив выполнение кластерного анализа, получим окно *Метод кластеризации* (рисунок 7.11). В этом окне задаются методы кластеризации в зависимости от того, известно ли заранее количество групп или нет. В нашем случае выбираем метод *Joining (tree clustering).*

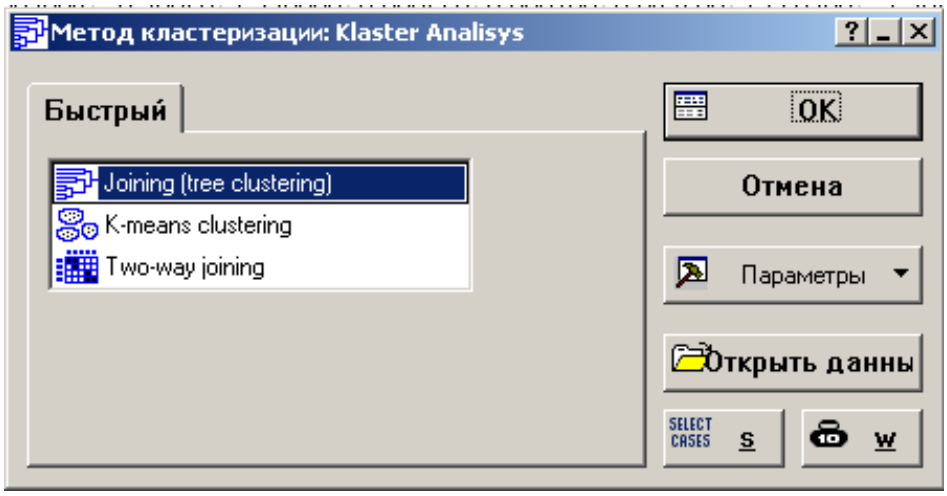


Рисунок 7.11 Выбор метода кластеризации

- Подтвердив выбранный метод, попадаем в окно *Кластерный анализ: Соединение* (рисунок 7.12). В этом окне войдите в раздел *Расширенный* и задайте исходные данные и способы выполнения анализа. В переменные *Variables* включите стандартизованные параметры инвестиционных фондов (колонку с названиями фондов включать не надо). Кластеры *Cluster* задаются по рядам *Cases (rows)*, так как группировать нужно фонды, которые в таблице расположены по строкам (рядам). В качестве *Правила объединения* выбираем распространенный *Ward's method*, а критерием классификации (*Измерение*) – *Squared Euclidean distance*.

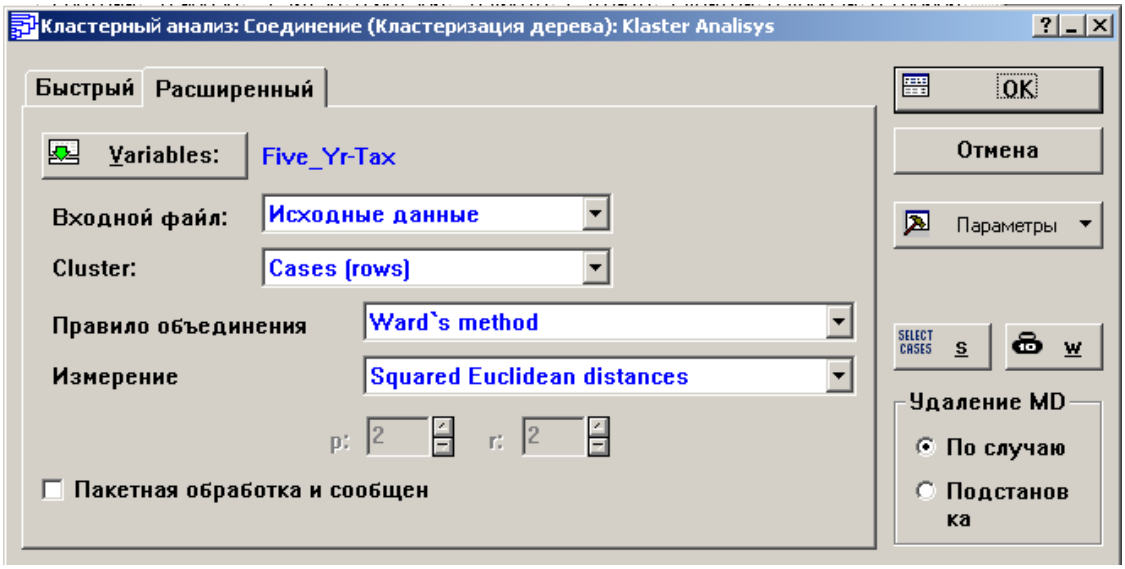


Рисунок 7.12 Задание исходных данных и способов выполнения анализа

- Запустив собственно выполнение кластерного анализа, получим окно *Результаты соединения* (рисунок 7.13).

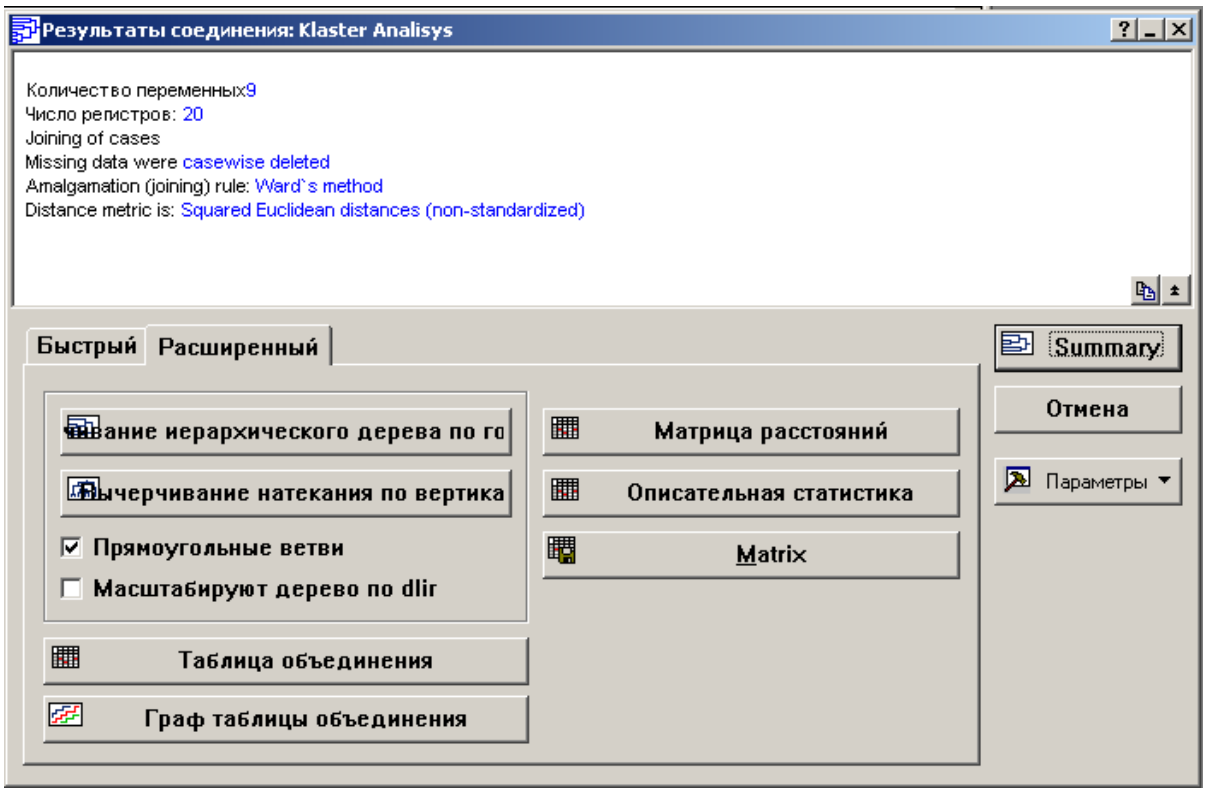


Рисунок 7.13 Окно *Результаты соединения*

- В окно *Результаты соединения* можно задать конечный график дерева кластеризации в горизонтальном (рисунок 7.14) или в вертикальном (рисунок 7.15) виде.

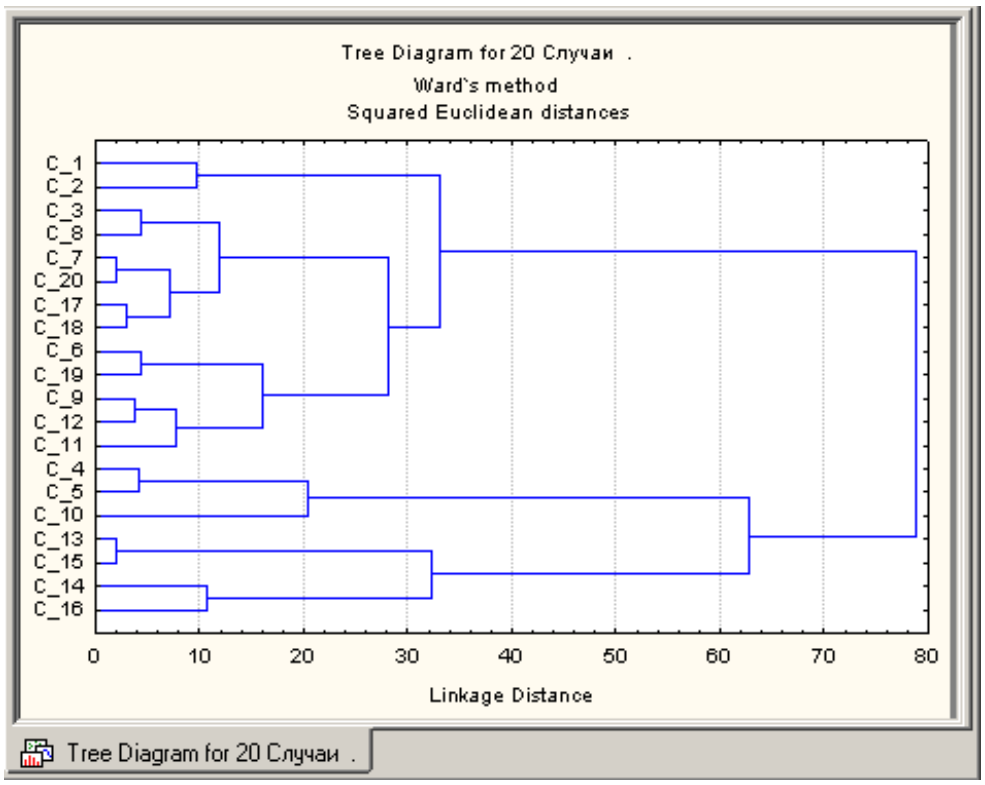


Рисунок 7.14 Горизонтальное дерево кластеризации

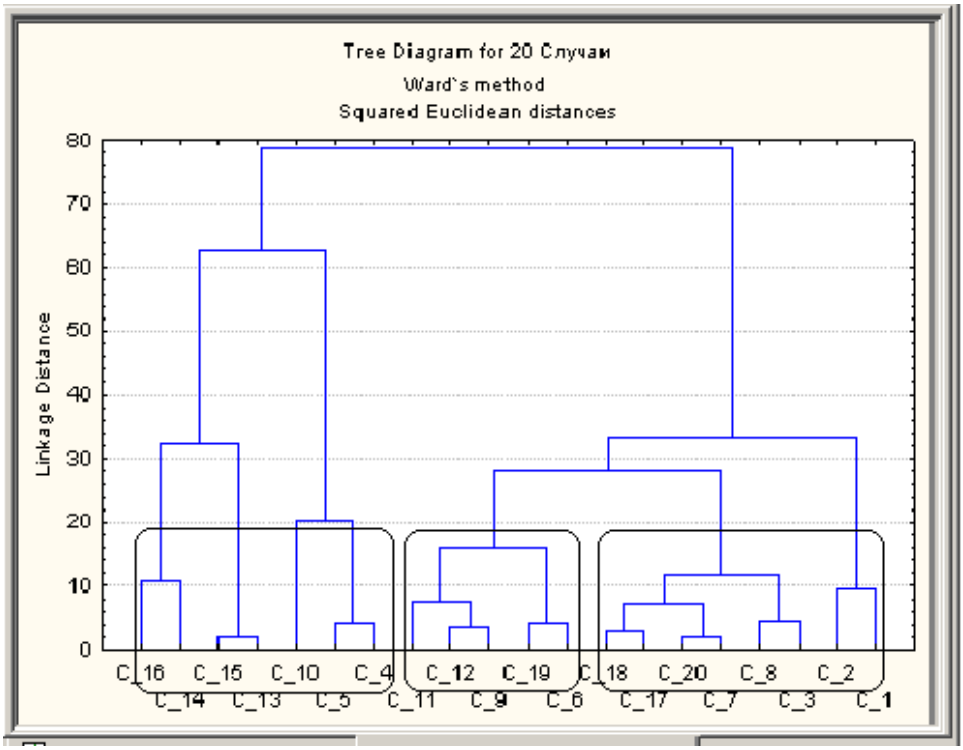


Рисунок 7.15 Вертикальное дерево кластеризации

- Так как нам необходимо разделить все фонды на три группы по степени надежности, разделим ветви дерева на три кластера (рисунок 7.16). На рисунке 7.16 показана результирующая таблица инвестиционных фондов с рекомендациями: надежные – Buy, ненадежные – Sell, середнячки – Hold .

Fund	Five Yr	Risk	Perf 94	Perf 93	Perf 92	Perf 91	Perf 90	Epe-nse	Tax	Rec-om-mend
F. Chip	16476	2	10	25	6	55	4	1.22	89	Buy
F. Contra	15476	2	-1	21	16	55	4	1.03	90	Buy
F. Destiny	14757	3	4	26	15	39	-3	0.7	69	Buy
Vista A	15145	4	-1	20	13	71	-6	1.49	96	Hold
Berger 100	15596	5	-7	21	9	89	-6	1.7	95	Hold
Gab. Assett	13640	1	0	22	15	18	-6	1.33	85	Buy
Neub. Focus	14081	3	1	16	21	25	-6	0.85	75	Buy
F. Magellan	13827	3	-2	25	7	41	-5	0.96	73	Buy
Janus	13187	2	-1	11	7	43	-1	0.91	85	Sell
L. Mason Value	13029	4	1	12	11	35	-17	1.82	92	Hold
Gabelli Growth	12301	3	-3	11	4	34	-2	1.41	80	Buy
Franklin Growth	11793	2	3	7	3	27	2	0.77	90	Sell
Janus 20	12441	4	-7	3	2	69	1	1.02	95	Sell
AARP Capital	11728	4	-10	16	5	41	-16	0.97	68	Sell
Kemper Growth	11386	4	-6	2	-2	67	4	1.09	86	Sell
20thCent.Growth	11258	4	-8	15	-4	32	0	1	60	Buy

Рисунок 7.16 Результирующая таблица инвестиционных фондов с рекомендациями

8 К какой группе относится объект? (Постановка диагноза)

8.1 Основы дискриминантного анализа

В предыдущем разделе рассматривался кластерный анализ, с помощью которого можно было сгруппировать объекты. Дискриминантный анализ тоже предусматривает отнесение объектов к той или иной группе. В чем же заключается отличие *кластерного* и *дискриминантного анализов*?

Кластерный анализ предназначен для того, чтобы сгруппировать объекты в однородные группы (кластеры). Эта однородность определяется на основе признаков (факторов), которые рассматриваются в качестве параметров кластерного анализа. Число групп заранее не известно. Нет результативного признака или зависимой переменной.

Дискриминантный анализ действует несколько иначе. Вводится некая «зависимая» переменная, на основе которой эксперт выносит суждение о принадлежности исследуемого объекта к той или иной группе. По результатам наблюдений за объектом строятся линейные классификайионные модели, позволяющие рассчитывать «зависимую» переменную.

Например, имеется три уровня приверженности потребителей к определенной марке товара и есть измерения ряда показателей жизни потребителя. Строятся линейные классификационные модели, соответствующие каждой марке товара, рассчитываются значения «зависимой» переменной по каждой модели, и определяется приверженность потребителя к определенной марке товара по максимальному значению «зависимой» переменной.

Выражаясь медицинскими терминами, потребителю ставится диагноз.

При выполнении процедуры дискриминации желательно, чтобы вероятность неправильного вывода в среднем была минимальной. Поясним это требование на примере двух множеств данных (рисунок 8.1).

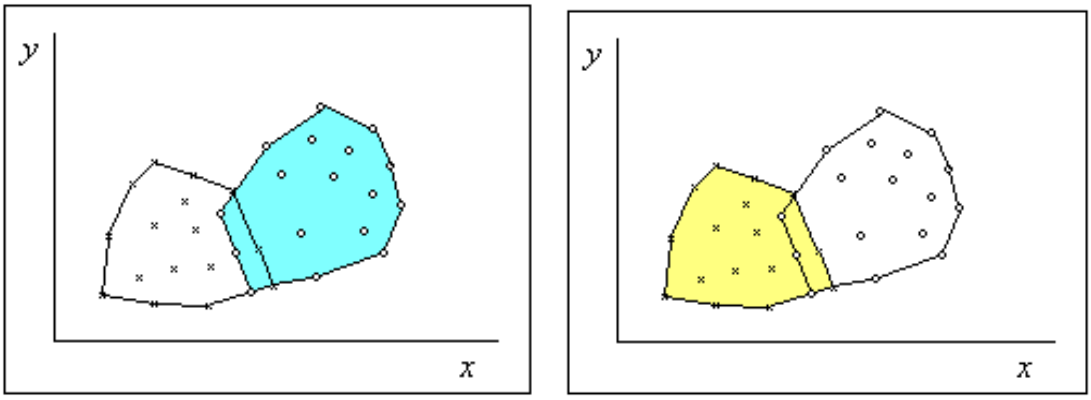


Рисунок 8.1 Варианты отнесения объектов к разным множествам

Два множества разделены, но в то же время имеют и пересечения, иначе бы не было проблемы дискриминации. Возможно появление ошибок, т.е. объект может быть отнесен не к своему множеству.

Дискриминантный анализ направлен на нахождение решения, удовлетворяющего следующему критерию:

$$J = \frac{\text{Дисперсия между классами}}{\text{Дисперсия внутри классов}} = \max.$$

8.2 Порядок выполнения

Исследуемые объекты разделяются экспертом на группы (классы). Проверка правильности распределения объектов по классам проводится на основании метрики Махаланобиса. Для этого с помощью этой метрики определяется расстояние от всех n объектов до центра тяжести каждой группы, определяемой по обучающей выборке. Отнесение экспертом i-го объекта в j-ю группу считается ошибочным, если расстояние Махаланобиса

от объекта до центра его группы значительно выше, чем от него до центра других групп, а вероятность попадания в свою группу ниже критического значения. В этом случае объект считается некорректно отнесенным и должен быть перемещен в другую группу.

Если таких объектов несколько, то сначала перемещается тот объект, у которого наибольшее расстояние Махаланобиса. При перемещении объекта из группы смещаются центры тяжести этой группы и той группы, в которую помещен перемещаемый объект. Поэтому возможно появление новых некорректных оценок. Перемещать нужно за один раз только один объект.

Процедура повторяется до тех пор, пока общий коэффициент корректности достигнет 100%, т.е. все объекты помещены в те группы, в которых соблюдаются требования метрики Махаланобиса.

Дискриминантный анализ позволяет рассчитать значения «зависимых» переменных – классификационных функций. При корректном распределении объектов по группам значения классификационных функций будут максимальными. Это означает, что для новых объектов можно рассчитать значения классификационных функций, и на основании этих расчетов поместить новые объекты в соответствующие группы, т.е. поставить диагнозы для этих объектов.

8.3 Указания к выполнению

Рассмотрим последовательность выполнения дискриминантного анализа на следующем примере. Исследованы 20 сельскохозяйственных предприятий, имеющих такие показатели работы:

X_1 – прибыль (тыс. р.);

X_2 – валовая продукция на 1 работника (тыс. р.);

X_3 –.валовая продукция на 1 га сельхозугодий (тыс. р.);

X_4 – производство молока на 1 га сельхозугодий (кг);

X_5 – производство мяса на 1 га сельхозугодий (кг);

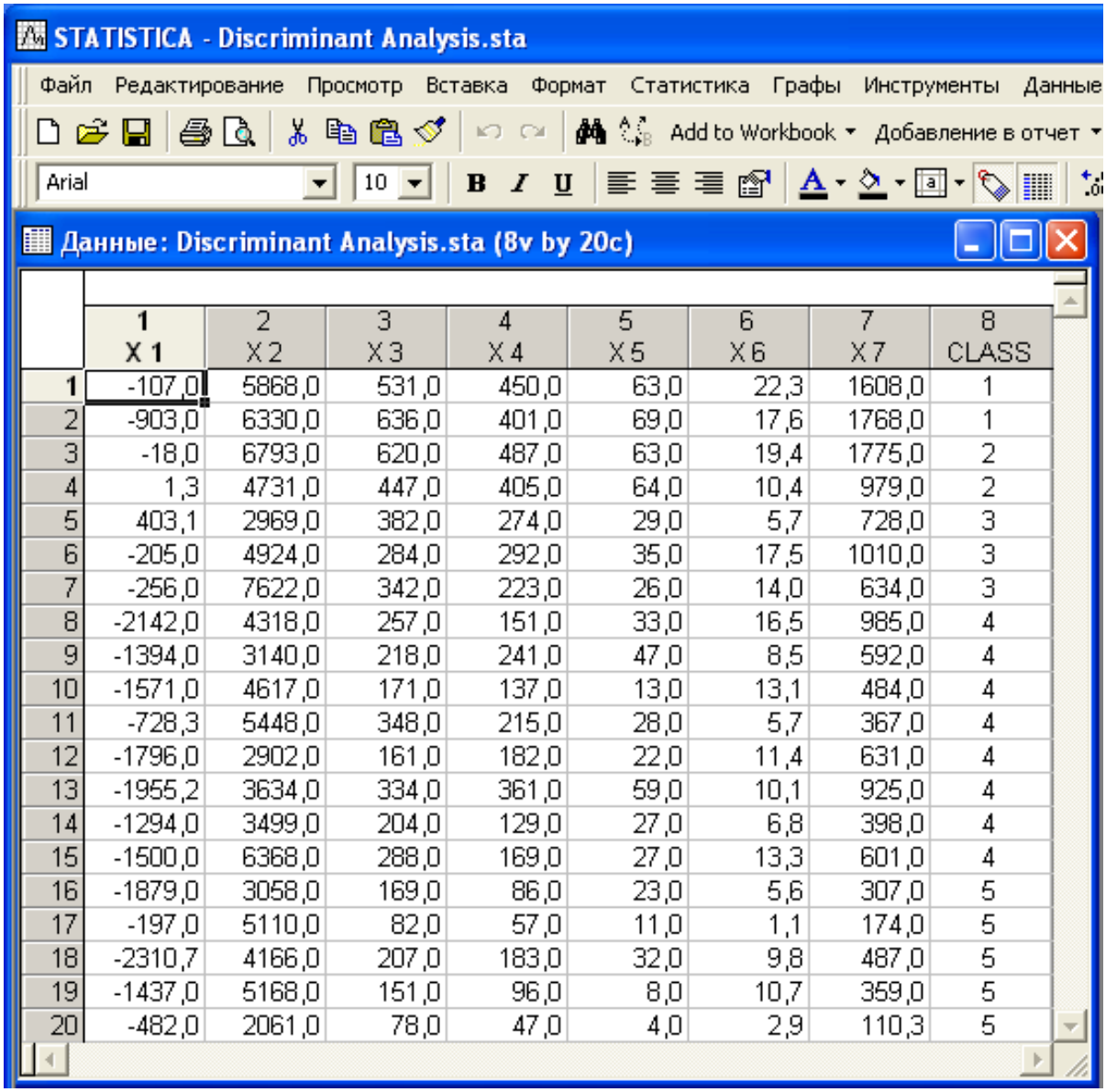
X_6 – выручка от реализации продукции на 1 работника (тыс. р.);

X_7 – выручка на 1 га сельхозугодий (тыс. р.).

Эксперты разделили исследованные объекты на 5 классов. Проверим достоверность такой классификации.

Для проведения дискриминантного анализа выполним следующие действия:

- В пакете ***STATISTICA for WINDOWS*** создадим базу данных (рисунок 8.2).



	1 X1	2 X2	3 X3	4 X4	5 X5	6 X6	7 X7	8 CLASS
1	-107,0	5868,0	531,0	450,0	63,0	22,3	1608,0	1
2	-903,0	6330,0	636,0	401,0	69,0	17,6	1768,0	1
3	-18,0	6793,0	620,0	487,0	63,0	19,4	1775,0	2
4	1,3	4731,0	447,0	405,0	64,0	10,4	979,0	2
5	403,1	2969,0	382,0	274,0	29,0	5,7	728,0	3
6	-205,0	4924,0	284,0	292,0	35,0	17,5	1010,0	3
7	-256,0	7622,0	342,0	223,0	26,0	14,0	634,0	3
8	-2142,0	4318,0	257,0	151,0	33,0	16,5	985,0	4
9	-1394,0	3140,0	218,0	241,0	47,0	8,5	592,0	4
10	-1571,0	4617,0	171,0	137,0	13,0	13,1	484,0	4
11	-728,3	5448,0	348,0	215,0	28,0	5,7	367,0	4
12	-1796,0	2902,0	161,0	182,0	22,0	11,4	631,0	4
13	-1955,2	3634,0	334,0	361,0	59,0	10,1	925,0	4
14	-1294,0	3499,0	204,0	129,0	27,0	6,8	398,0	4
15	-1500,0	6368,0	288,0	169,0	27,0	13,3	601,0	4
16	-1879,0	3058,0	169,0	86,0	23,0	5,6	307,0	5
17	-197,0	5110,0	82,0	57,0	11,0	1,1	174,0	5
18	-2310,7	4166,0	207,0	183,0	32,0	9,8	487,0	5
19	-1437,0	5168,0	151,0	96,0	8,0	10,7	359,0	5
20	-482,0	2061,0	78,0	47,0	4,0	2,9	110,3	5

Рисунок 8.2 База данных сельскохозяйственных предприятий

- Для запуска дискриминантного анализа выполним следующую последовательность действий: *Статистика/Многомерные исследующие методы/Дискриминантный анализ* (рисунок 8.3).

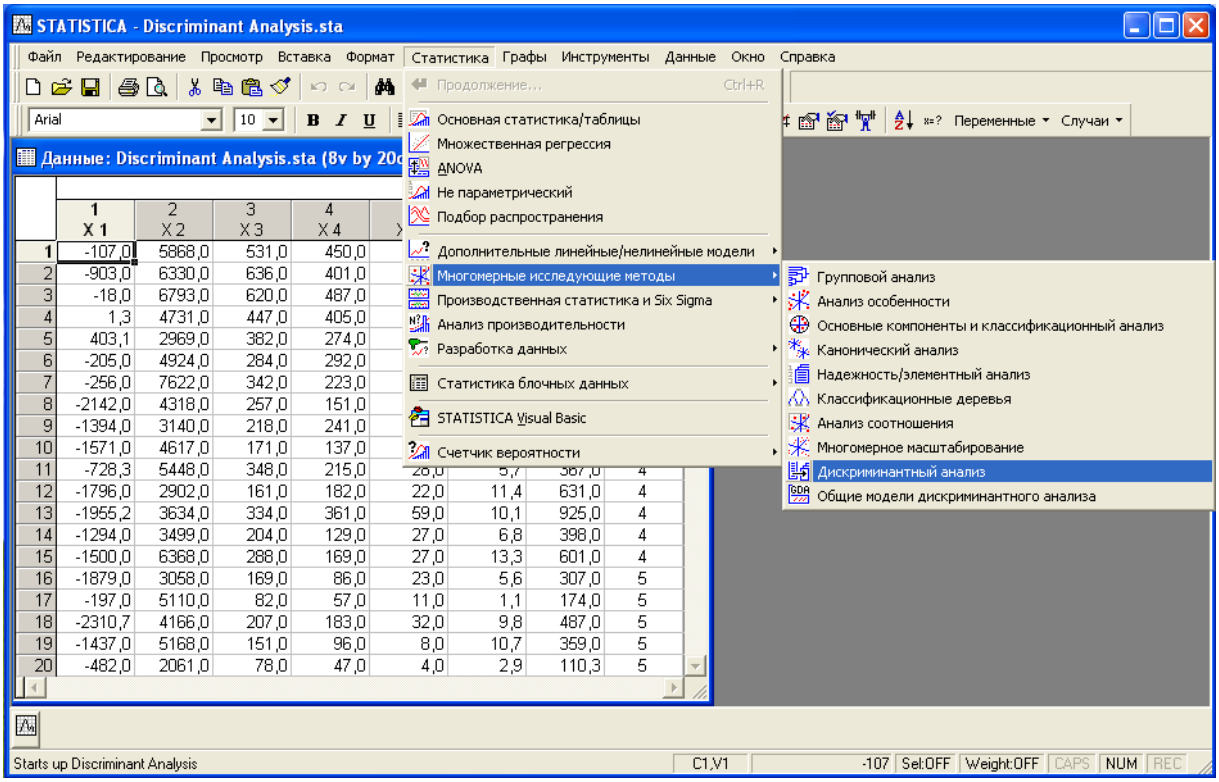


Рисунок 8.3 Запуск дискриминантного анализа

- Запустив выполнение дискриминантного анализа, получим окно *Дискриминантный анализ функции* (рисунок 8.4).

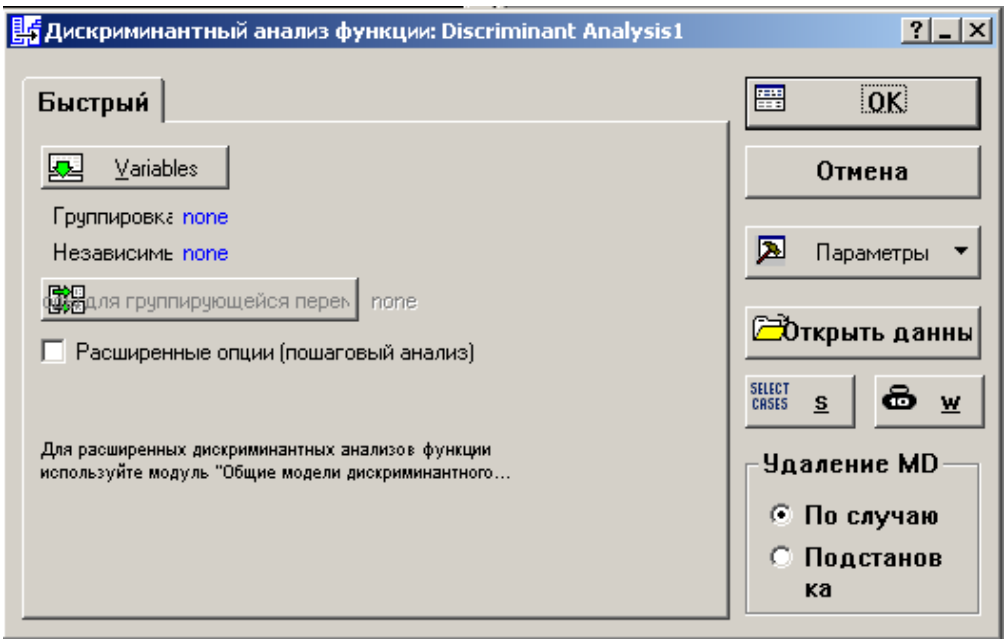


Рисунок 8.4 Окно *Дискриминантный анализ функции*

- В этом окне (рисунок 8.4) нужно задать переменные, для чего щелкните по клавише *Variables*. Получим окно выбора «зависимой» и независимых переменных (рисунок 8.5).

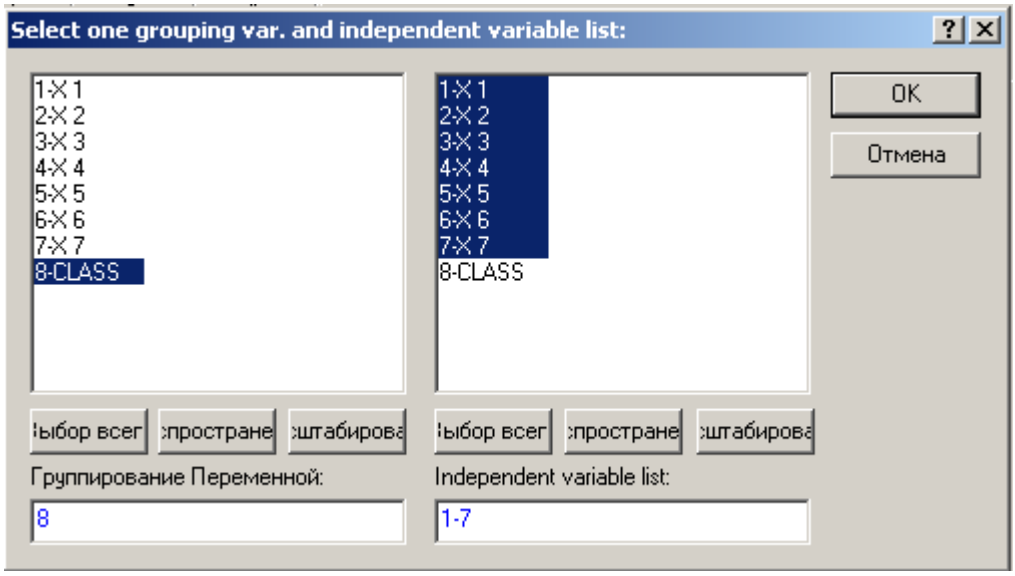


Рисунок 8.5 Окно выбора переменных

- По условиям задачи независимыми переменными являются показатели работы сельскохозяйственных предприятий X_1 – X_7. «Зависимой» переменной является класс предприятий *CLASS*, т.е. группы, к которым эксперты отнесли то или иное предприятие. Поэтому в окне (рисунок 8.4) слева зададим колонку 8 – (*CLASS)*, а справа: колонки 1-7 (X_1 – X_7). В результате вернемся в окно *Дискриминантный анализ функции* (рисунок 8.6), но уже с заданными переменными.

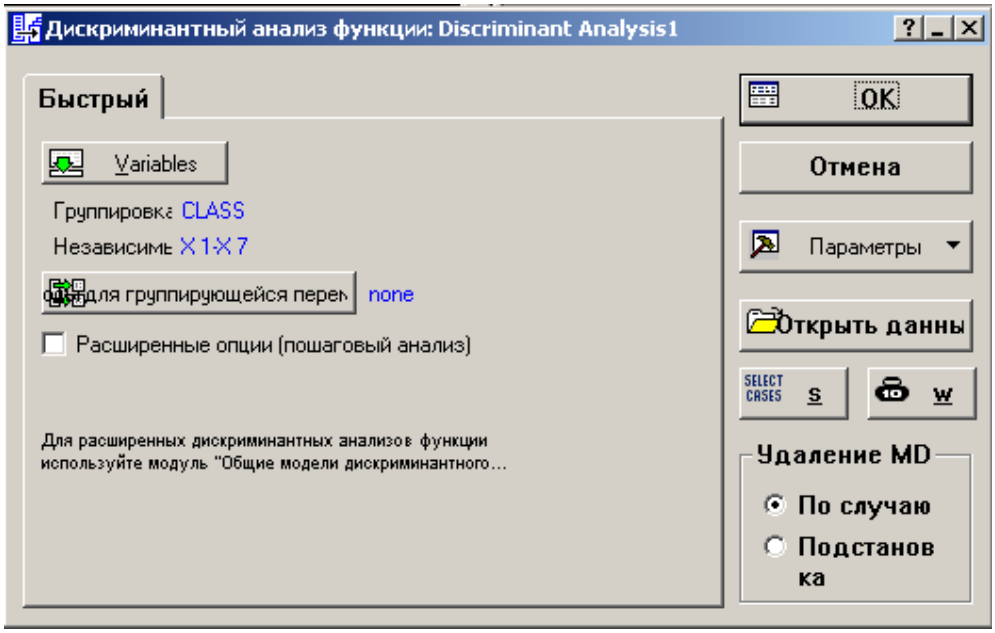


Рисунок 8.6 Окно *Дискриминантный анализ функции* с заданными переменными

- После задания переменных функций запустим выполнение расчета в окне *Дискриминантный анализ функции* (рисунок 8.6). Откроется окно *результатов дискриминантного анализа* (рисунок 8.7).

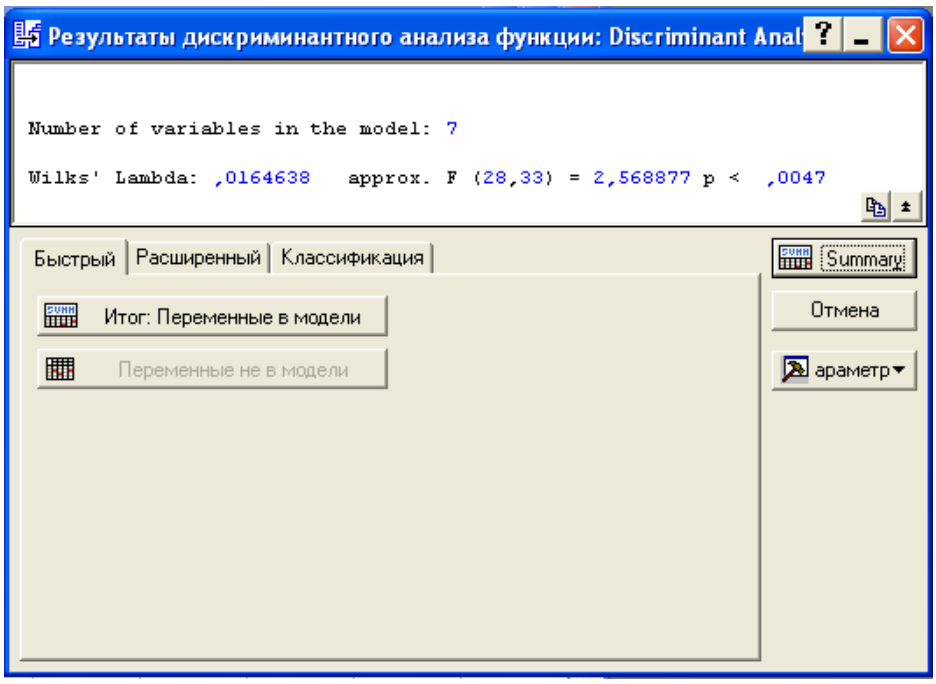


Рисунок 8.7 Окно *результатов дискриминантного анализа*

- В окне (рисунок 8.7) щелкнем на клавишу *Классификация*. В результате это окно примет вид (рисунок 8.8).

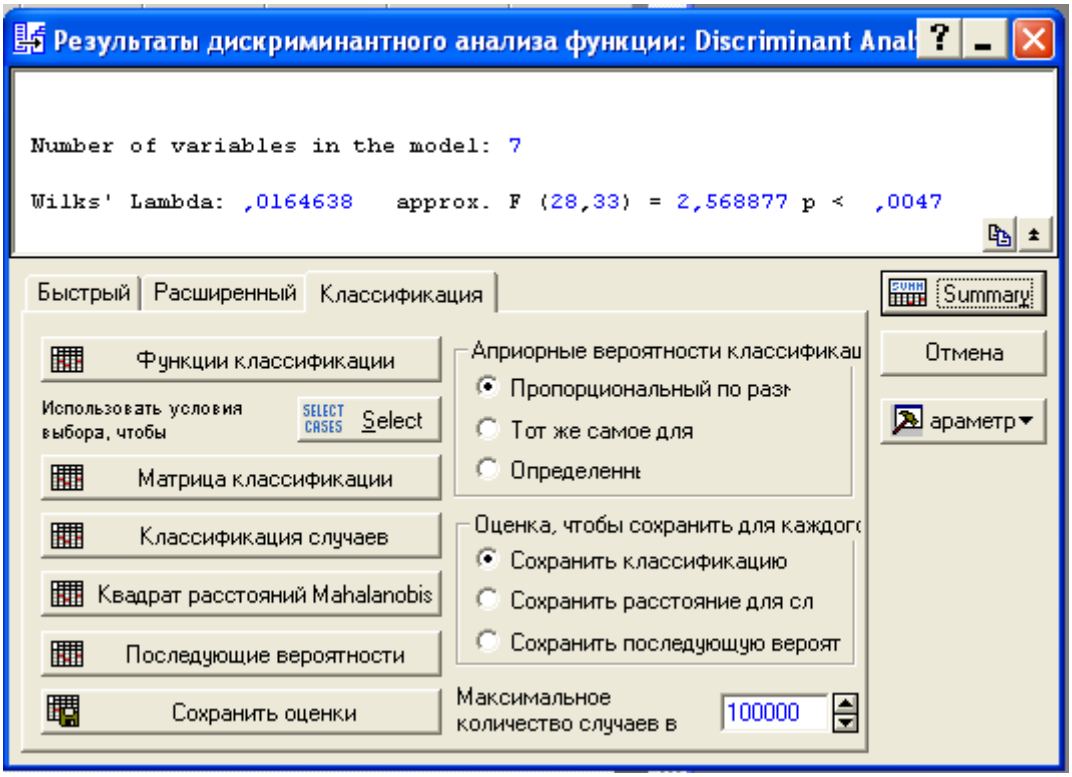


Рисунок 8.8 Окно *результатов классификации дискриминантного анализа*

- Информационная часть диалогового окна результатов дискриминантного анализа сообщает, что:

– Number of variables in the model (число переменных в модели) = 7;

– Wilks lambda (значение лямбды Уилкса) = 0,0164638;

– Approx. F (28,33) (приближенное значение F – статистики, связанной с лямбда Уилкса) = 2, 568867;

– p - уровень значимости < 0,0047.

Значение лямбды Уилкса лежит в интервале [0,1]. Так как это значение близко 0, это свидетельствует о хорошей дискриминации. Если бы значение лямбды Уилкса приближалось к 1, это свидетельствовало бы о плохой дискриминации. По данным лямбды Уилкса и значению p- уровня значимости, намного меньше допустимого 0,05, можно сделать вывод, что данная классификация практически корректна.

Для проверки корректности распределения всех сельскохозяйственных объектов по классам посмотрим результаты классификационной матрицы, для чего в окне (Рис. 8.8) щелкнем по клавише *Матрица классификаций*. В результате получим искомую матрицу (рисунок 8.9).

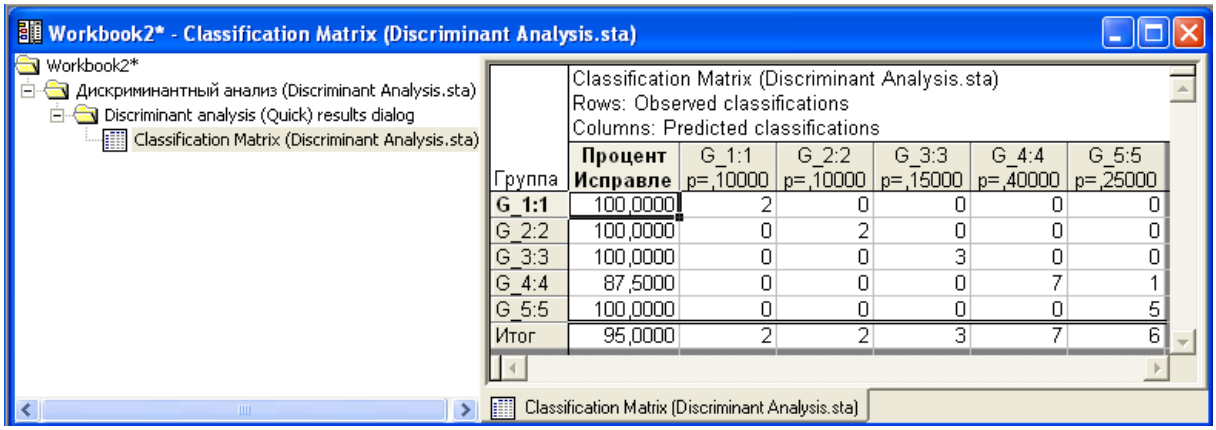


Рисунок 8.9 Окно *Матрица классификаций*

- В нашем примере в целом результаты классификации неплохие. Только в 4 классе не все результаты имеют 100 % достоверность. Т.е. могут быть отдельные неверно проклассифицированные предприятия. Для выяснения этого посмотрим результаты классификации, для чего в окне (Рис. 8.8) щелкнем на клавишу *Классификация случаев*. В результате появится соответствующее окно (рисунок 8.10).

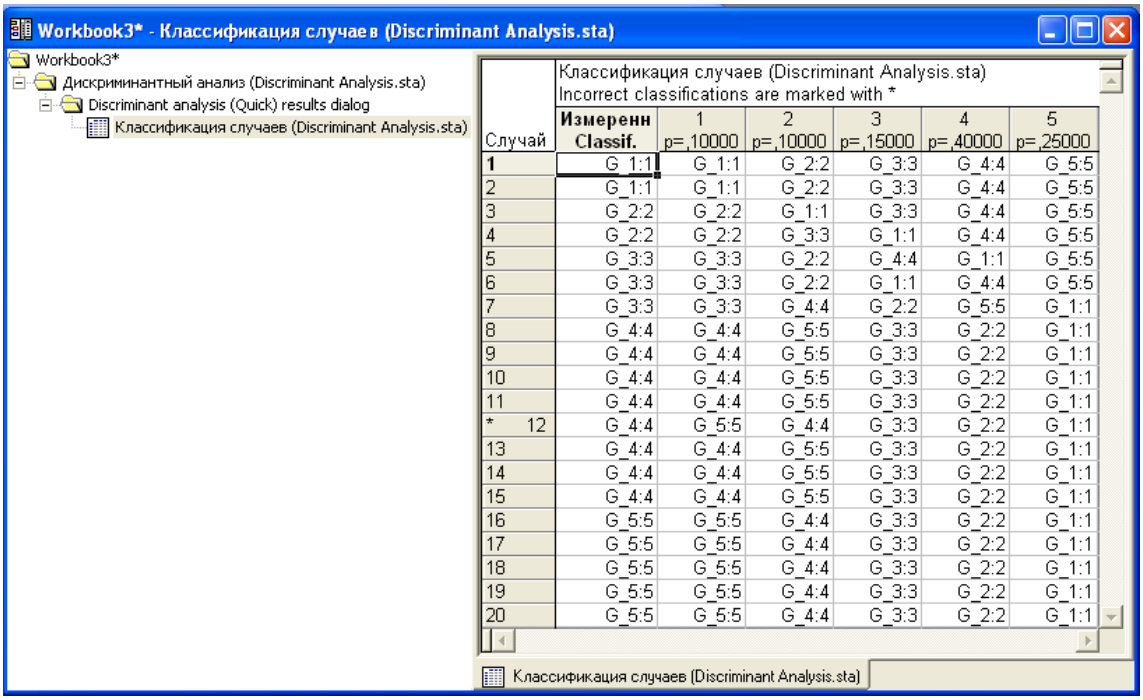


Рисунок 8.10 Окно *Классификация случаев*

- В окне *Классификация случаев* (рисунок 8.10) случаи некорректной классификации (Incorrect classifications are marked with *) отмечены звездочкой *. В нашем примере предприятие 12 неправильно классифицировано. Для исключения некорректных объектов с помощью метрики Махаланобиса определим расстояние от всех n объектов до центров тяжести каждой группы. Как отмечалось ранее, отнесение i-го объекта в j-ю группу считается ошибочным, если

расстояние Махаланобиса от объекта до центра его группы выше, чем от него до центра других групп. Для определения метрики Махаланобиса в окне результатов классификации дискриминантного анализа (рисунок 8.8) щелкнем по клавише *Квадрат расстояний Mahalanobis*. Появится окно (рисунок 8.11).

В этом окне видно, что для предприятия 12 квадрат расстояний Махаланобиса для четвертого класса, куда эксперт поместил предприятие 12, выше, чем для пятого класса.

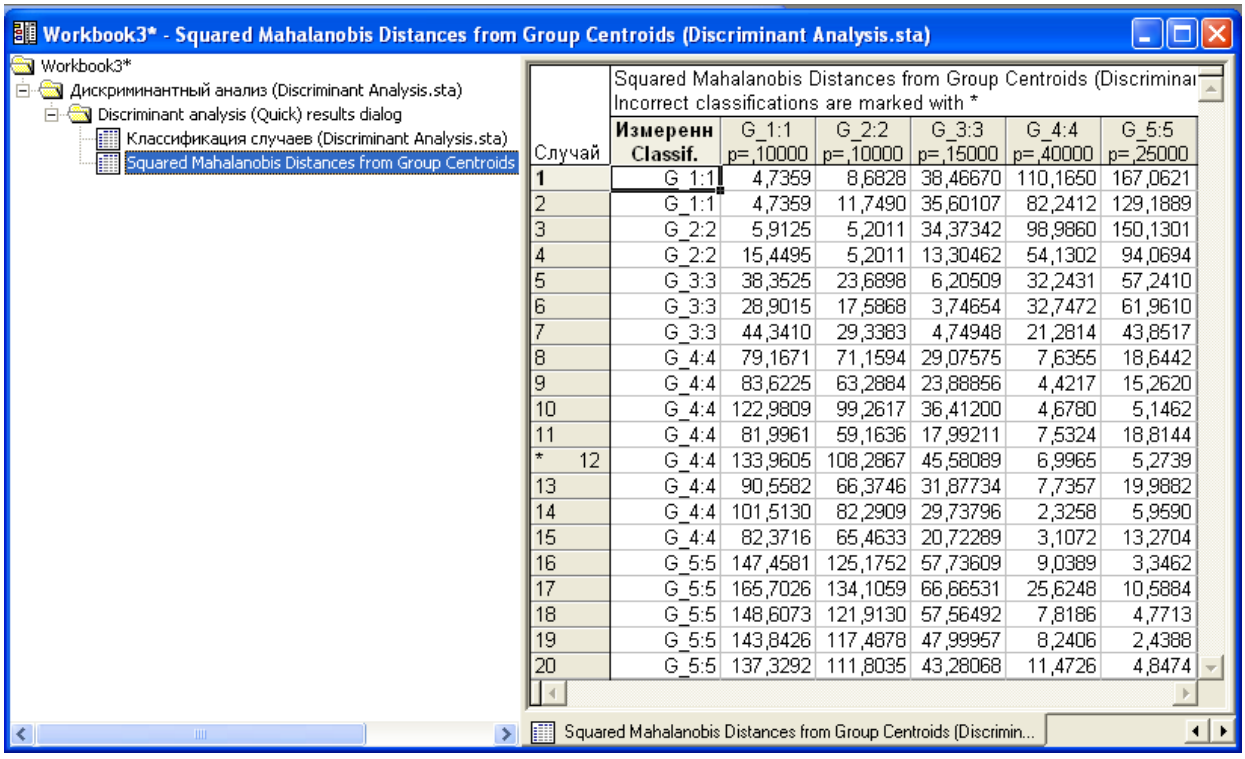


Рисунок 8.11 Окно *Квадрат расстояний Mahalanobis*

- Чтобы выправить ситуацию, нужно в таблице исходных данных заменить классификацию у предприятия 12: поставить вместо класса 4 класс 5. После чего снова провести всю процедуру дискриминантного анализа. Но на этом процедура может не закончиться. Дело в том, что после перемещения одного из объектов в другой класс происходит смещение центров тяжести классов. В

нашем примере так и произошло: после перемещения 12 предприятия в 5 класс некорректным стало помещение также 10 предприятия в 4 класс. После оценок метрики Махаланобиса появляется необходимость перемещения в 5 класс также 10 предприятия (рисунок 8.12).

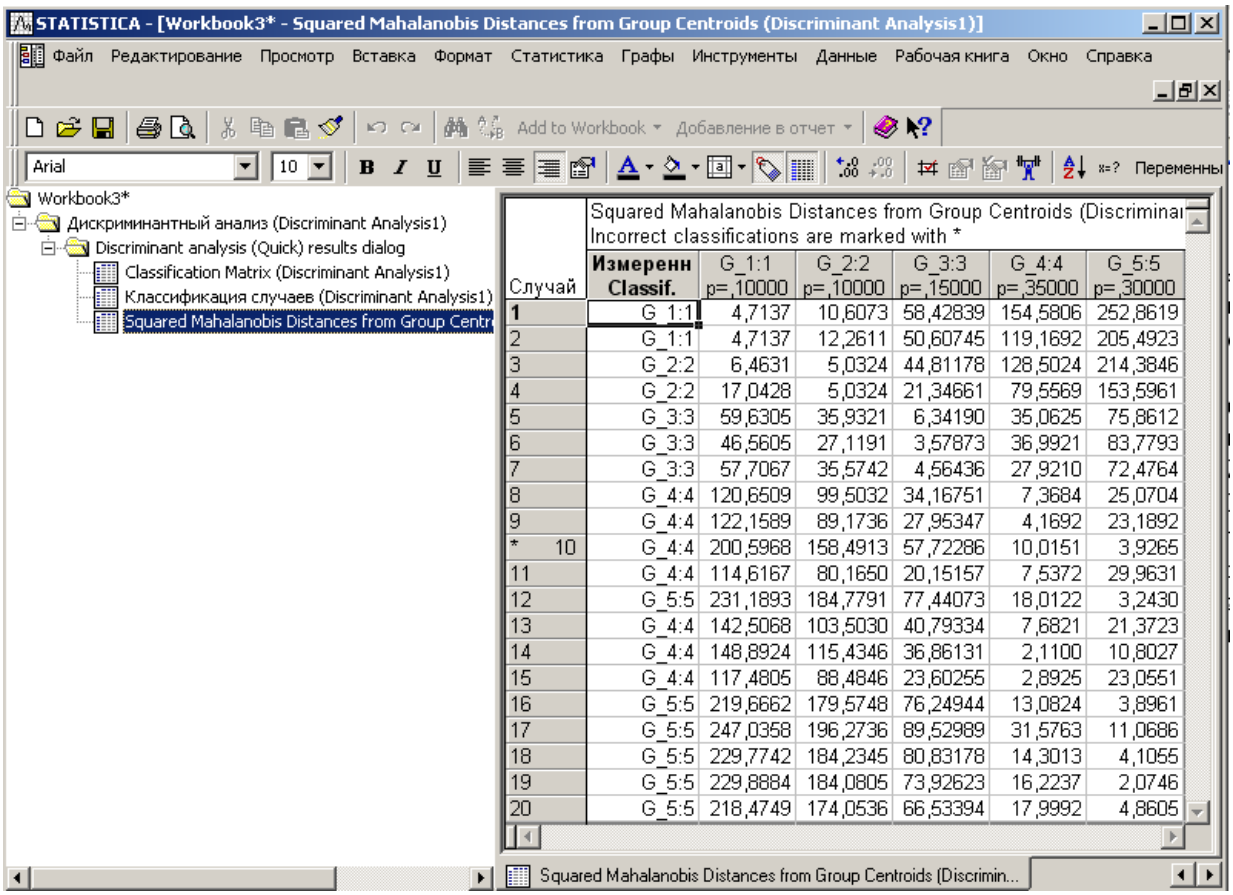


Рисунок 8.12 Окно *Квадрат расстояний Mahalanobis* после первой корректировки

- Вернемся опять в таблицу исходных данных и заменим классификацию у предприятия 10: поставим вместо класса 4 класс 5, так как квадрат расстояний Махаланобиса для пятого класса у этого предприятия самый маленький.
- Для проверки корректности всех случаев откроем снова Окно *Матрица классификаций* (рисунок 8.13).

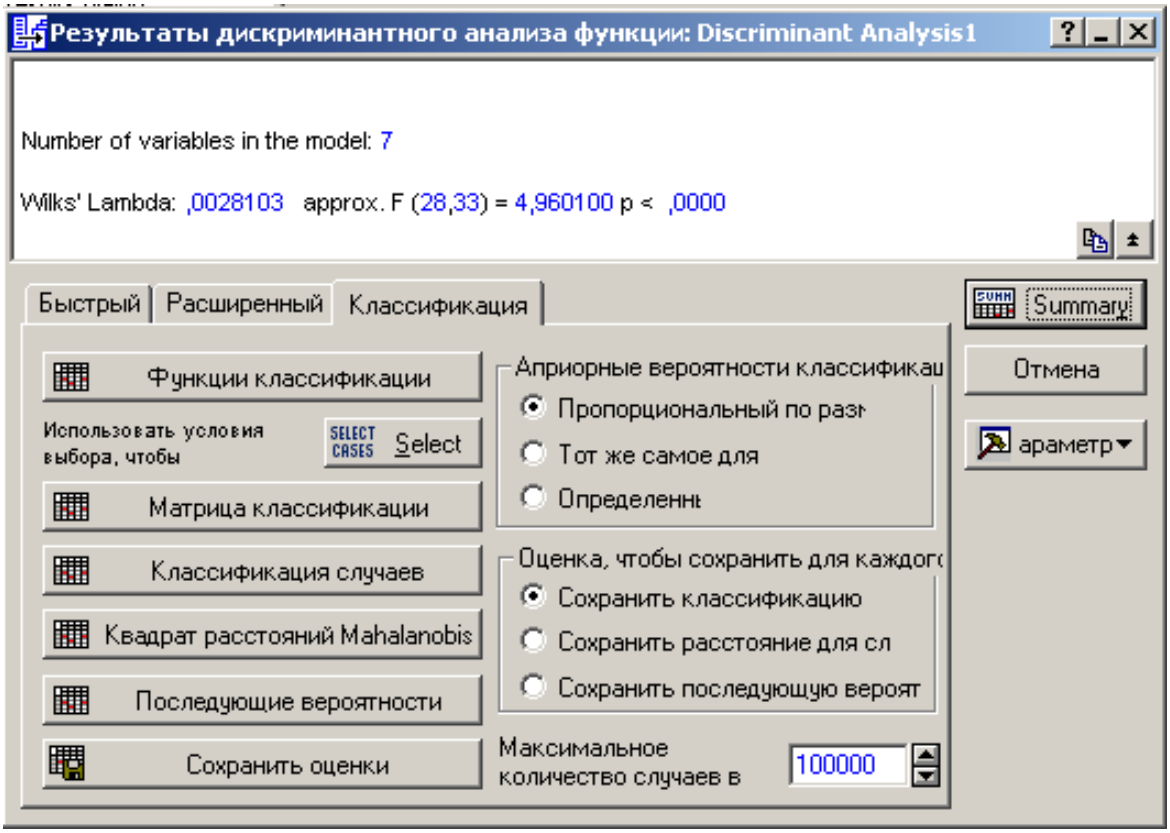


Рисунок 8.13 Окно *результатов классификации дискриминантного анализа* после вторичной корректировки

- Теперь все в порядке: все группы (классы) имеют 100% достоверность классификации (рисунок 8.14).

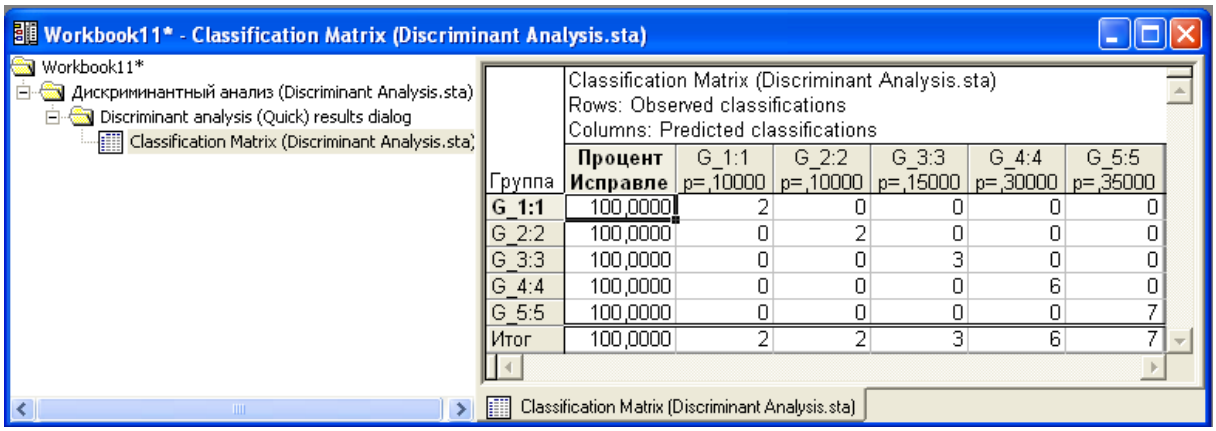


Рис. 8.14. Окно *Матрица классификаций* после корректировок.

- На рисунке 8.14. показаны откорректированная *база данных* и *метрики Махаланобиса* после окончательной корректировки.

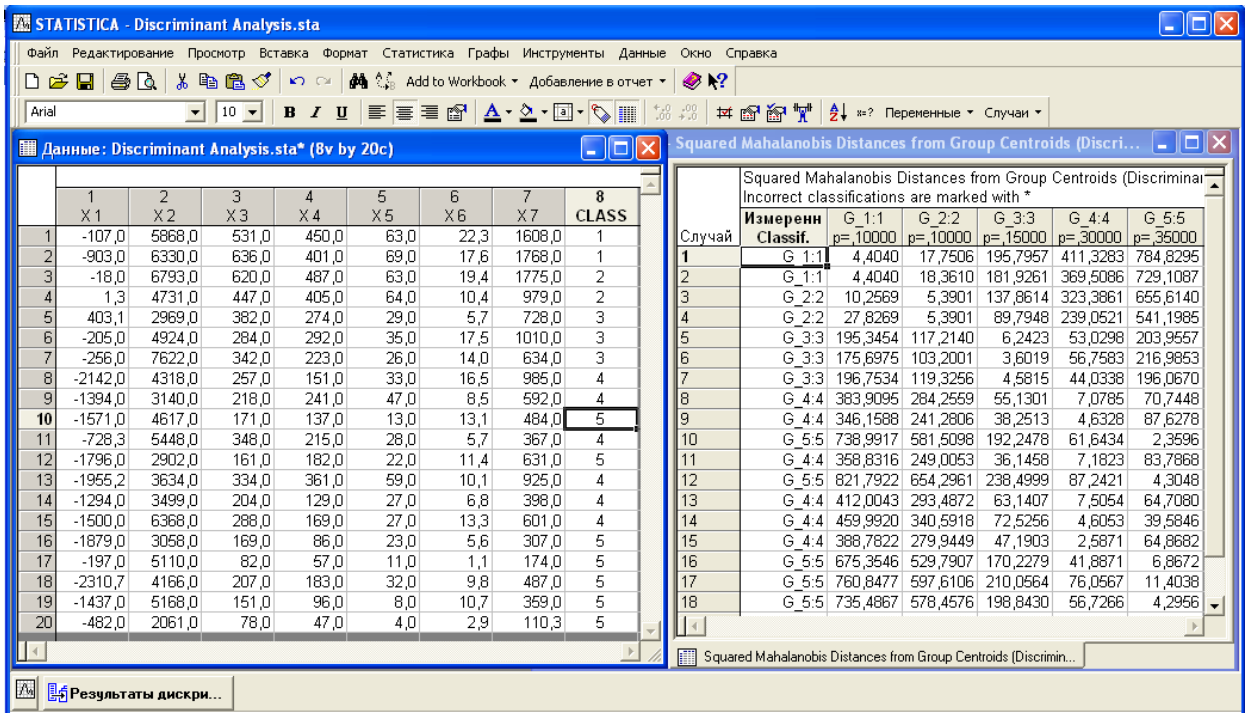


Рисунок 8.14. Откорректированная *база данных* и *метрики Махаланобиса*.

- Теперь можно вычислить классификационные функции. Для этого в окне результатов классификации дискриминантного анализа (рисунок 8.13) щелкните по клавише *Функции классификации*. Появится соответствующее окно (рисунок 8.15).

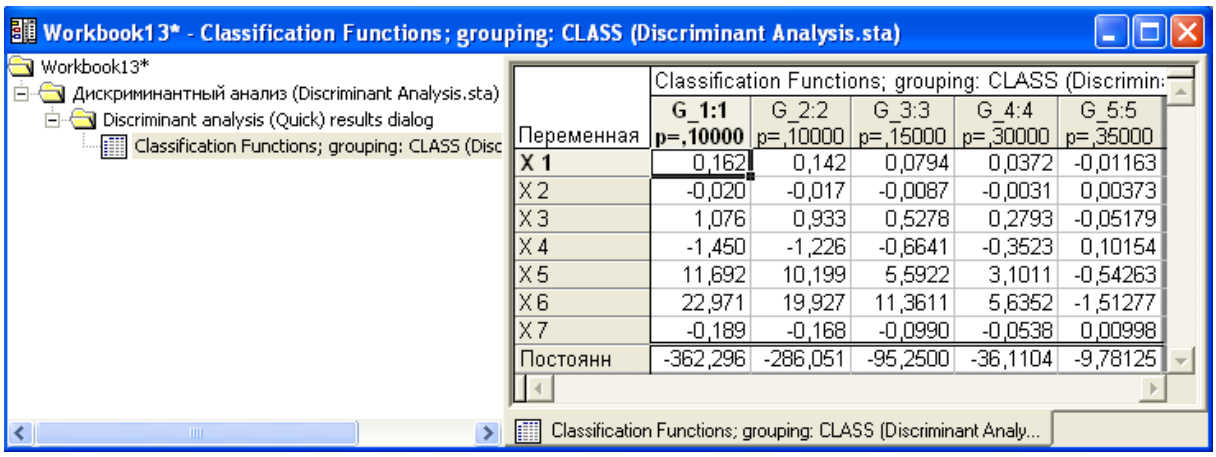


Рисунок 8.15 Окно *Функции классификации*

- Из окна *Функции классификации* можно составить классификационные функции для каждого класса. Например, для первых двух классов классификационные функции будут выглядеть следующим образом:

  Y_1 = -326,296 + 0,162X1 – 0,020X2 + 1,076X3 – 1,450X4 + 11,692X5 + 22,971X6 – 0,189X7;

  Y_2 = -286,051 + 0,142X1 – 0б017X2 + 0,933X3 – 1,226X4 + 10,199X5 + 19,927X6 – 0,168X7.

  Напомним, что при корректном распределении объектов по группам классификационные функции будут иметь максимальные значения.

9 Эффективны ли нововведения?

9.1 Основы однофакторного дисперсионного анализа

При исследовании зависимостей одной из наиболее простых является ситуация, когда можно указать только один фактор, влияющий на конечный результат, и этот фактор может принимать лишь конечное число значений (уровней). Такие задачи (называемые задачами дисперсионного однофакторного анализа) весьма часто встречаются на практике. Типичный пример задач однофакторного дисперсионного анализа – сравнение по результатам нескольких различных способов действия, направленных на достижение одной цели, скажем, повышение зарплаты или нескольких видов рекламы и т.д.

Данные.

Для сравнения влияния факторов на результат необходим определенный статистический материал. Обычно его получают следующим образом: каждый из k уровней фактора применяют несколько раз (не обязательно одно и то же число раз) к исследуемому объекту и регистрируют результаты. Например, исследуется результативность продаж с различным видом рекламы или производительность труда рабочих, с различной оплатой труда и т.д. Итогом подобных испытаний являются k выборок.

Наиболее распространенным и удобным способом представления подобных данных является таблица (см. таблицу 9.1).

Таблица 9.1 База данных для исследования влияния фактора

Уровни фактора	1	2	...	k	...	m
Результаты измерений	x_{11}	x_{12}	...	x_{1k}	...	x_{1m}
	x_{21}	x_{22}	...	x_{2k}	...	x_{2m}
	.	.	...	.	...	.
	x_{i1}	x_{i2}	...	x_{ik}	...	x_{im}
	.	.	...	.	...	.
	$x_{n1}1$	$x_{n2}2$	...	$x_{nk}k$	...	x_{nkm}

Здесь m – число уровней исследуемого фактора, n_j – число наблюдений *над k-м уровнем фактора,* $n_1, \ldots, n_m$ - объемы выборок с различными уровнями исследуемого фактора, $N = n_1 + n_2 + \ldots + n_m$ – общее число наблюдений.

Каждое наблюдение можно представить суммой вида:

x_{ik} – *i – элемент ,* $i = 1,\ldots, n;$ *k* -выборки $k = 1, \ldots m.$

Основная идея метода заключается в изучении источников изменчивости зависимой переменной (отклика) и разложения общей дисперсии наблюдаемых значений отклика на составляющие – дисперсию, обусловленную влиянием изучаемого фактора, и остаточную дисперсию, являющуюся следствием действия случайных причин и неучтенных факторов. Сравнивая дисперсию факторов с остаточной дисперсией, можно проверить гипотезу о влиянии фактора.

Среднее значение k-выборки (9.1)

$$\overline{x}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ik}$$

Общее среднее всех наблюдений $\overline{x} = \frac{1}{n}\sum_{k=1}^{m}\sum_{i=1}^{n_k} x_{ik}.$ (9.2)

Основное тождество дисперсионного анализа имеет следующий вид:

$$Q = Q_1 + Q_2, \qquad (9.3)$$

где Q – общая сумма квадратов отклонений (дисперсий) наблюдаемых значений x_{ik} от общего среднего $\overline{x}$; Q_1 – сумма квадратов отклонений выборочных средних $\bar{x}_k$ от общего среднего $\overline{x}$ (сумма квадратов отклонений между группами); Q_2 – сумма квадратов отклонений наблюдаемых значений x_{ik} от выборочной средней $\bar{x}_k$ (сумма квадратов отклонений внутри групп).

Расчет этих сумм квадратов отклонений осуществляется по следующим формулам:

$$Q = \sum_{k=1}^{m} \sum_{i=1}^{n_k} (x_{ik} - \bar{x})^2 = \sum_{k=1}^{m} \sum_{i=1}^{n_k} x_{ik}^2 - n\,\bar{x}^2 \qquad (9.4)$$

$$Q_1 = \sum_{k=1}^{m} n_k (\bar{x}_k - \bar{x})^2 = \sum_{k=1}^{m} n_k \bar{x}_k^2 - n\,\bar{x}^2 \qquad (9.5)$$

$$Q_2 = \sum_{k=1}^{m} \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k)^2 = \sum_{k=1}^{m} \sum_{i=1}^{n_k} x_{ik}^2 - \sum_{k=1}^{m} n_k \bar{x}_k^2 \qquad (9.6)$$

9.2 Порядок выполнения

Прежде чем судить о количественном влиянии фактора на измеряемый признак, полезно спросить себя, есть ли такое влияние вообще. Нельзя ли объяснить расхождения наблюденных в опыте значений для разных уровней фактора действием чистой случайности? Ведь внутренне присущая явлению изменчивость уже привела к тому, что результаты оказываются различными даже при неизменном значении фактора (т.е. внутри каждого столбца табл. 9.1). Может быть, той же причиной можно объяснить и различие между ее столбцами? На статистическом языке это предположение означает, что все данные табл. 9.1 принадлежат одному и тому же распределению. Это

предположение обычно именуют нулевой гипотезой и обозначают $\mathbf{H}_0$. Т.е. исследуемый фактор не оказывает значимого влияния. Если оно оказывается справедливым, то анализ заканчивается. В противном случае возникает задача оценки величины эффектов обработки и выяснения качества полученных оценок.

Метод носит название *однофакторного дисперсионного анализа.* Это название связано с тем, что анализ модели (8.3) основан на сопоставлении двух оценок дисперсии σ^2. Одна дисперсия связана с результаты действия исследуемых факторов, а другая – с независимыми случайными величинами, отражающими внутреннюю изменчивость.

Если верна гипотеза $\mathbf{H}_0$: $\sigma^2_1 = \ldots = \sigma^2_m$, то это означает, что исследуемый фактор не оказывает значимого влияния. Если гипотеза не верна, то исследуемый фактор, например, нововведения действительно оказывает влияние.

Таким образом, однофакторный дисперсионный анализ сводится к сравнению двух дисперсий: от исследуемого фактора и от внутренней нестабильности. Оценка дисперсий выполняется с помощью ***F***-распределения *Фишера*:

$$F = \frac{Q_1/(m-1)}{Q_2/(n-m)} \qquad (9.7)$$

Если вычисляемое ***F*** - *отношение* оказывается больше табличного при выбранном уровне значимости (степени риска), например ά = 0,05, то нулевая гипотеза $\mathbf{H}_0$ отвергается, т.е. исследуемый фактор оказывает влияние.

Пример.

Три группы торговцев продавали штучный товар, расфасованный в различные упаковки. После окончания распродажи был произведен тестовый контроль над случайно отобранными продавцами из каждой группы. Результаты контроля представлены в таблице 9.2.

Таблица 9.2 Результаты продаж товаров с разными упаковками

Номер группы	1	2	3
Число продаж продавцов, x_{ik}	1 3 2 1 0 2 1	2 3 2 1 4	4 5 3
Общее число продаж	10	12	12
Количество продавцов	7	5	3

В данном примере число выборок m = 3, число продаж во всех выборках n = 15, отсюда:

$\bar{x}_1$ = 10/7 = 1,428;

$\bar{x}_2$ = 12/5 = 2,4;

$\bar{x}_3$ = 12/3 = 4;

$\bar{x}$ = (1,428 + 2,4 + 4)/3 = 2,226.

Если $\sum_{k=1}^{m} \sum_{i=1}^{n_k} x_{ik}^2$ = 1 + 9 + 4 + 1 + 4 + 1 + 4 + 9 + 1 + 16 + 16 + 25 + 9 = 104,

$\sum_{k=1}^{m} n_k \bar{x}_k^2 = 7*1{,}428^2 + 5*2{,}4^2 + 3*4^2 = 91{,}074$,

тогда: $Q = 104 - 15*\ 2{,}226^2 = 26{,}93$,

$Q_1 = 91{,}074 - 15*2{,}226^2 = 14{,}01$,

$Q_2 = Q - Q_1 = 26{,}93 - 14{,}01 = 12{,}92$.

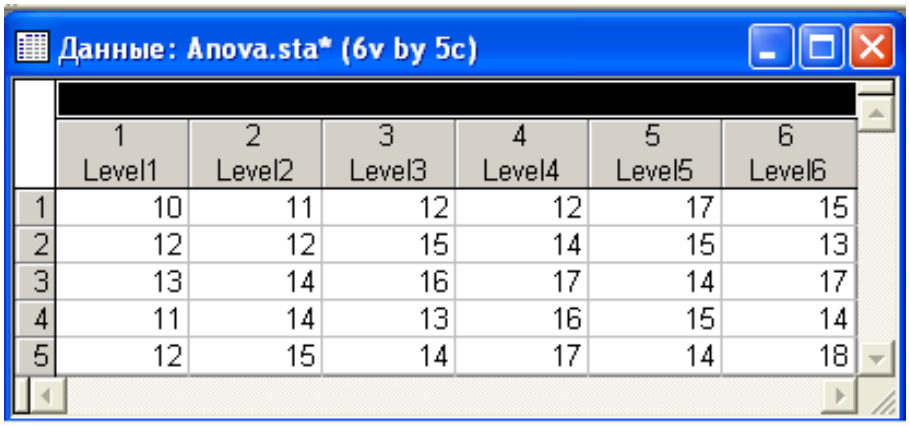
Вычислим критерий Фишера

$$F = \frac{14,01/2}{12,92/12} = 6,52 .$$

Сравнивая это значение при уровне значимости ά = 0,05 с табличным $F > F_{2;12} = 3,885$, делаем вывод, что упаковка (особенно красочная!) влияет на количество продаж.

9.3 Указания к выполнению

- Рассмотрим выполнение однофакторного дисперсионного анализа для оценки нововведений в пакете ***STATISTICA for WINDOWS.***
- Проиллюстрируем применение описанных выше критериев на следующем примере. Для выяснения влияния денежного стимулирования на результаты труда шести однородным группам из пяти человек каждая были предложены задачи одинаковой трудности. Задачи предлагались каждому испытуемому независимо от всех остальных. Группы отличаются между собой величиной денежного вознаграждения за решаемую задачу. Составим таблицу (рисунок 9.1) с числом решенных задач членами каждой группы.



	1 Level1	2 Level2	3 Level3	4 Level4	5 Level5	6 Level6
1	10	11	12	12	17	15
2	12	12	15	14	15	13
3	13	14	16	17	14	17
4	11	14	13	16	15	14
5	12	15	14	17	14	18

Рисунок 9.1. Пример таблицы данных для однофакторного анализа.

В качестве нулевой гипотезы $\mathbf{H}_0$ примем предположение об отсутствии влияния денежного вознаграждения на число решенных задач и проверим эту гипотезу.

Откроем окно *ANOVA* (рисунок 9.2). Сокращение *ANOVA* происходит от выражения “Analysis of variance”. В отечественной литературе вместо термина “анализ вариации” чаще используется термин “дисперсионный анализ”.

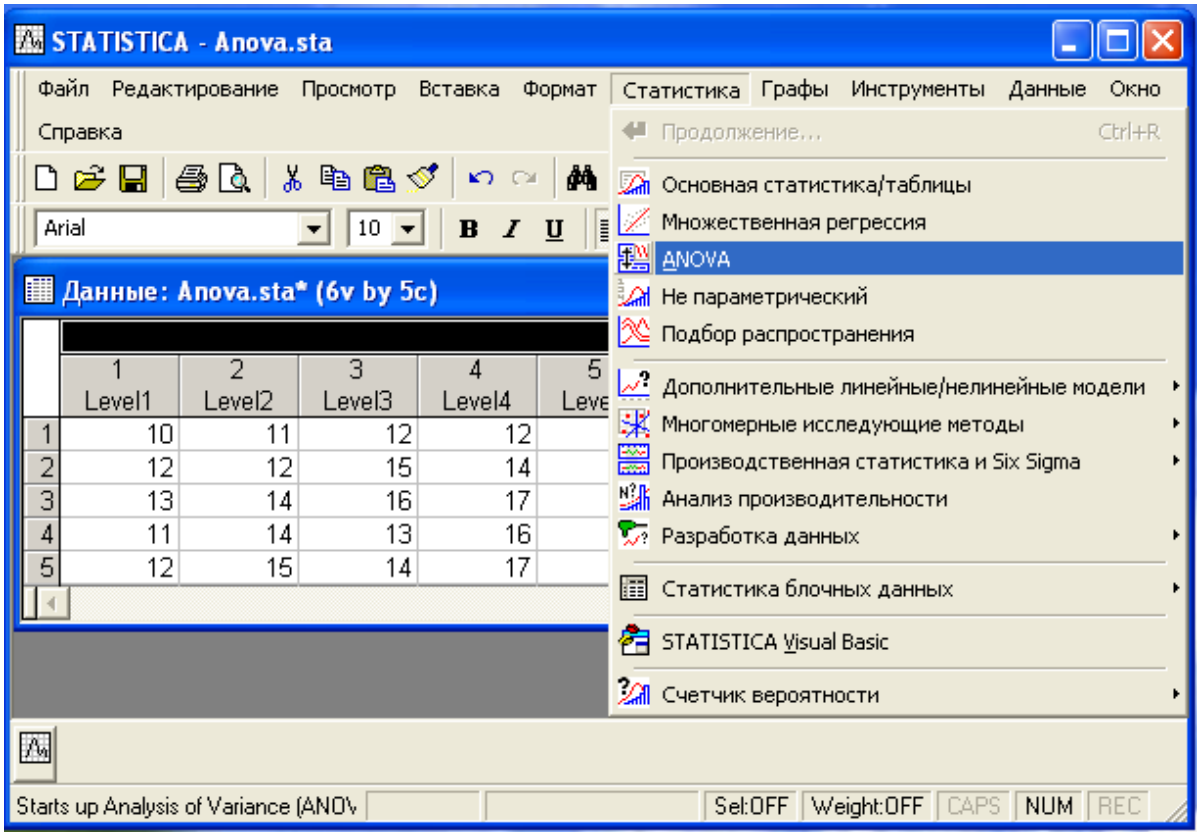


Рисунок 9.2 Окно *ANOVA*

- Запустив выполнение процедуры *ANOVA*, мы получим окно выбора типа анализа и метода спецификаций (рисунок 9.3).

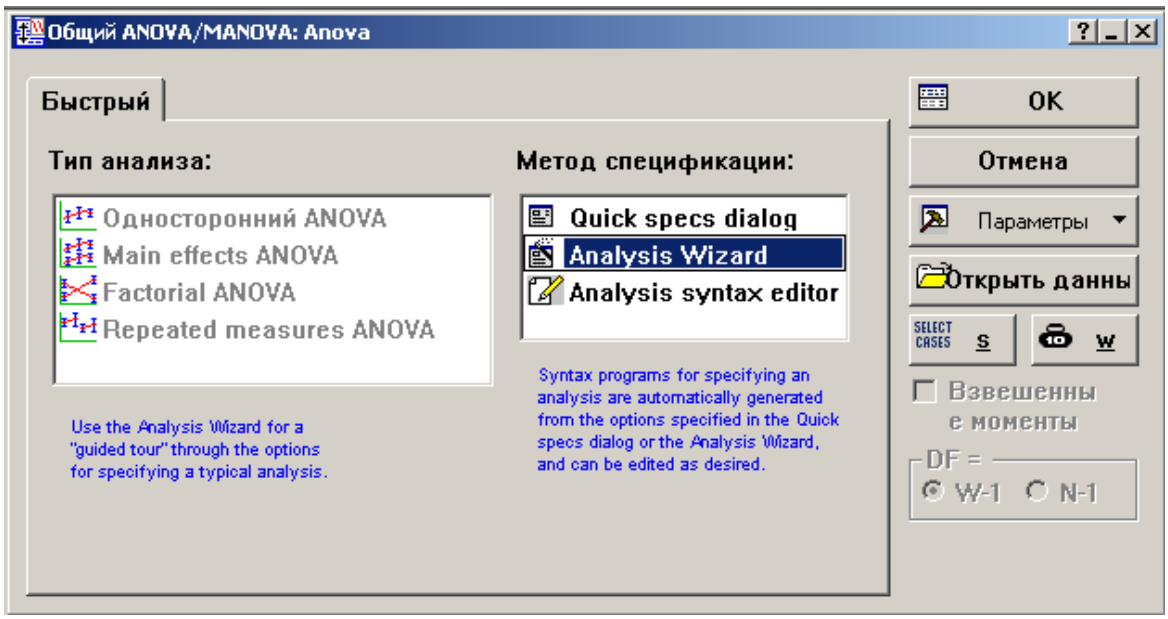


Рисунок 9.3 Окно выбора типа анализа и метода спецификаций

- В окне выбора типа анализа и метода спецификаций выберем в качестве метода спецификаций *Analysis Wizard*. Подтвердив этот метод, получаем окно *Внешнего расчета* (рисунок 9.4).

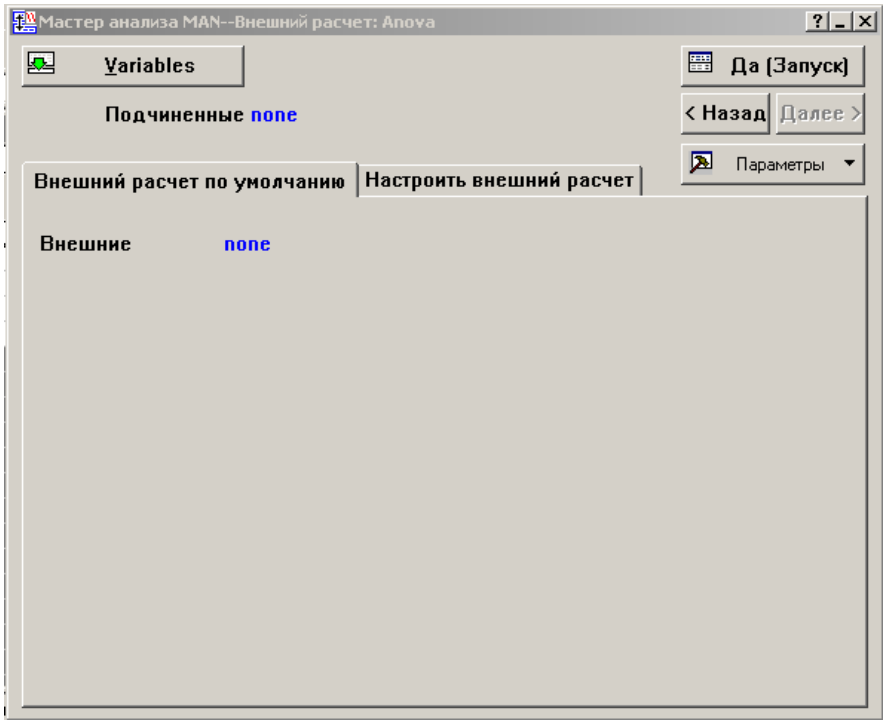


Рисунок 9.4 Окно *Внешнего расчета* без заданных переменных

- Следующим шагом является задание переменных. Для этого подтвердим запуск в окне *Внешнего расчета.* В результате откроется окно выбора переменных (рисунок 9.5).

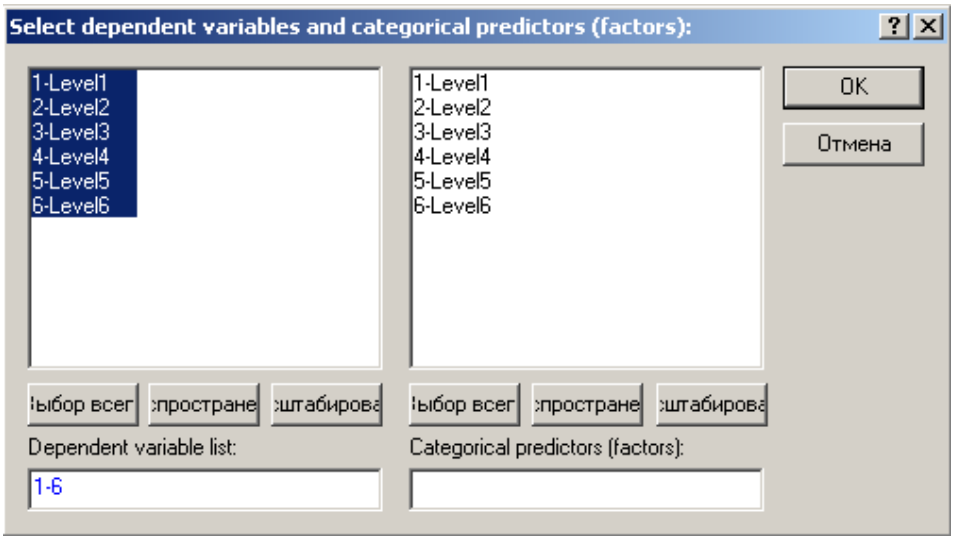


Рисунок 9.5 Окно выбора переменных

- Заполнив в качестве зависимых все переменные (независимых переменных нет), запустим выполнение. В результате получим Окно *Внешнего расчета* с заданными переменными (рисунок 9.6).

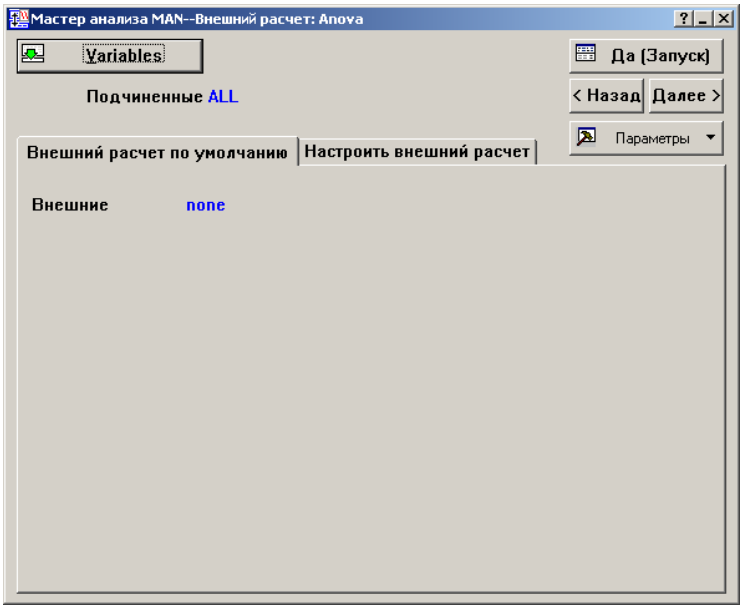


Рисунок 9.6 Окно *Внешнего расчета* с заданными переменными

- Подтвердив запуск внешнего расчета, получаем окно первых результатов *Result 1* (рисунок 9.7).

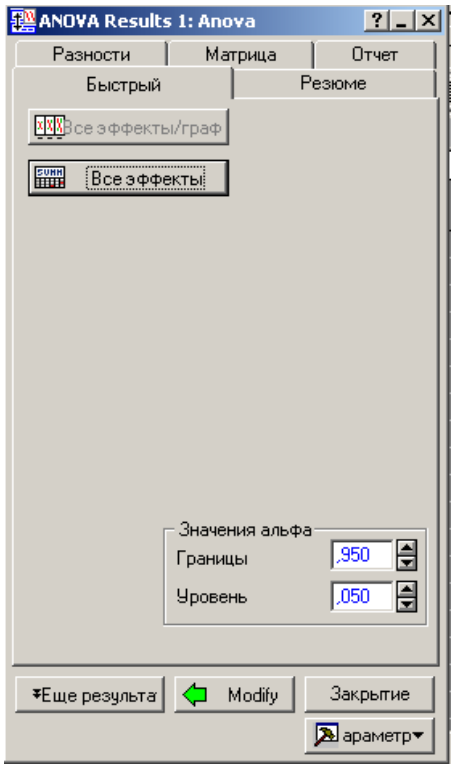


Рисунок 9.7 Окно первых результатов *Result 1*.

- В окне *Result 1* запускаем процедуру *Все эффекты*. После этого появляется Результирующая таблица однофакторного анализа (рисунок 9.8).

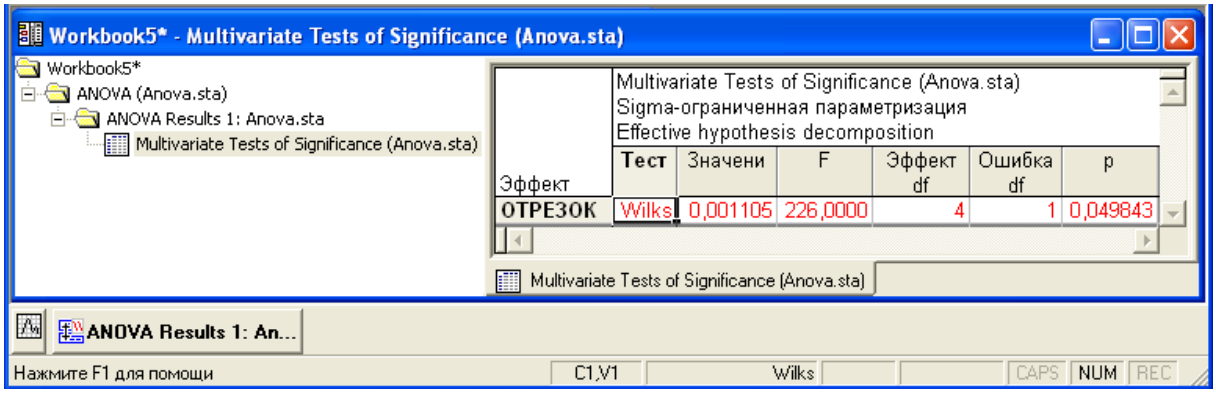


Рисунок 9.8 Результирующая таблица однофакторного анализа

Вместо вычисления ***F - отношений*** с заданным уровнем значимости и последующего сравнения с табличным значением пакет ***STATISTICA for WINDOWS*** вычисляет **p-level**. Величина **p-level** характеризует тот уровень значимости, при котором дисперсия между колонками отличается от дисперсии внутри колонок. Т.е., если **p-level** меньше 0,05, это означает, что для уровня значимости ά = 0,05 можно считать, что исследуемый фактор влияет на результирующую переменную.

В нашем примере **p-level** = 0, 049843, т.е. меньше 0,05. Поэтому можно считать, что исследуемый фактор (денежное стимулирование) влияет на результаты труда.

СПИСОК ЛИТЕРАТУРЫ

1. Боровиков В.П. STATISTICA: искусство анализа данных на компьютере. – СПб: Питер, 2004.

2. Грабауров В.А. Информационные технологии для менеджеров. – М.: Финансы и статистика, 2-е изд. 2005.

3. Дюк В. Обработка данных на ПК в примерах. СПб., 1997.

4. Лабоцкий В.В. Многомерная обработка экономических данных с использованием интегрированной системы STATISTICA. – Мн.: БГЭУ, 2002.

5. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. – М.: Инфра - М. Финансы и статистика. 2003.

Учебное издание

СТАТИСТИЧЕСКАЯ ОБРАБОТКА ИНФОРМАЦИИ С ПОМОЩЬЮ ПАКЕТА «STATISTICA»

Учебно-методическое пособие

Составитель
Грабауров Владимир Александрович

Ответственный за выпуск *В.А. Грабауров*

Издано в редакции автора

Подписано в печать 25.06.2008 г. Формат $60\times84^{1}/_{8}$
Бумага офсетная. Гарнитура Times New Roman. Усл. печ. л. 11,0.
Уч.-изд. л. 5,4. Тираж 120 экз. Заказ 592.

Издатель и полиграфическое исполнение
Белорусский государственный аграрный технический университет
ЛИ № 02330/0131734 от 10.02.2006. ЛП № 02330/0131656 от 02.02.2006.
220023, г. Минск, пр. Независимости, 99, к. 2.