МИНИСТЕРСТВО СЕЛЬСКОГО ХОЗЯЙСТВА И ПРОДОВОЛЬСТВИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Саратовский государственный аграрный
университет им. Н.И. Вавилова

ИСПОЛЬЗОВАНИЕ ПАКЕТА STATISTICA 5.0 ДЛЯ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ОПЫТНЫХ ДАННЫХ

Методические указания
для дипломного проектирования
для студентов лесного факультета специальностей
260400 "Лесное хозяйство" и
260500 "Садово-парковое и ландшафтное строительство"

Саратов 2001

Использование пакета Statistica 5.0 для статистической обработки опытных данных: Методические указания для дипломного проектирования для студентов лесного факультета специальностей 260400 "лесное хозяйство" и 260500 "садово-парковое и ландшафтное строительство"/ / Сост.: С.В. Кабанов. Сарат. гос. агр. ун-т. Саратов, 2000. с.

Рецензенты:

- доцент кафедры лесомелиорации СГАУ им. Н.И. Вавилова, к.с.-х.н. В.Н. Филатов;

- зав. кафедрой ботаники и экологии СГУ им. Н.Г. Чернышевского, профессор, д.б.н. В.А. Болдырев.

Методические указания составлены с учетом накопленного опыта проведения лабораторных занятий на кафедре и обобщения литературных данных. Были использованы также соответствующие пособия и методические указания, изданные кафедрами других вузов страны.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Методические указания предназначены для оказания помощи студенту по работе с программой Statistica по проведению статистического анализа данных. В первую очередь они будут полезны студентам-дипломникам, работающим над своими дипломными работами и проектами. Пакет Statistica занимает в мире устойчиво лидирующее положение среди программ статистической обработки данных. В последнее время появились первые подробные пособия (см. список литературы), посвященные работе с этим пакетом. Однако эта литература для массового пользователя не всегда является легко доступной

На простых примерах, касающихся различных сторон лесного дела, показаны возможности пакета по первичной обработке опытных данных и множественному регрессионному анализу. Методические указания рассматривают всего два (наиболее часто использующихся в опытном лесном деле) из большого количества статистических модулей, имеющихся в программе.

Методические указания рассчитаны на пользователя, имеющего начальные навыки работы с Windows-программами.

Составитель методических указаний выражает благодарность рецензентам за ценные советы и замечания.

ПЕРВИЧНАЯ ОБРАБОТКА ОПЫТНЫХ ДАННЫХ ПРИ ПОМОЩИ МОДУЛЯ Basic Statistics / Tables

Расчет описательных статистик производится при помощи модуля Basic Statistic/Tables. В этом модуле объединены наиболее часто использующиеся на начальном этапе обработки данных процедуры.

В стартовой панели модуля приводится перечень статистических процедур этого модуля (рис. 2):

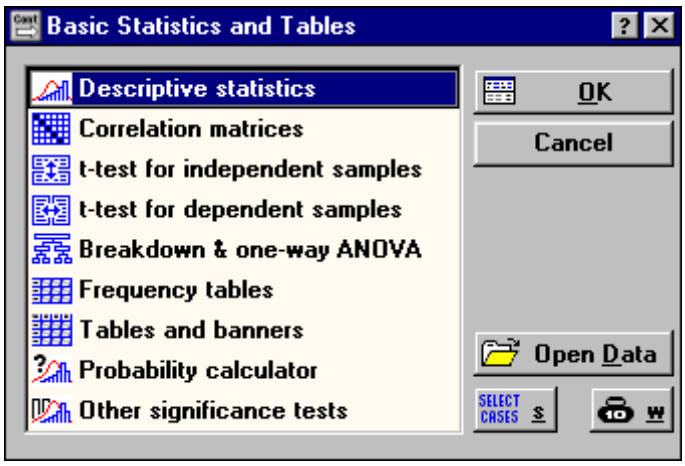


Рис. 2. Стартовое окно модуля с перечнем статистических процедур

Descriptive statistics - Описательные статистики;
Correlation matrices - Корреляционные матрицы;
t-test for independent samples - t-тест для независимых выборок;
t-test for dependent samples - t-тест для зависимых выборок;
Br eakdown = one-way ANOVA - Классификация и однофакторный дисперсионный анализ; и др.

2.1. Процедура Descriptive statistics (Описательные статистики)

Рассмотрим возможности этой процедуры на примере.

Имеется выборка объемом 50 измерений, представляющая собой результаты обмера 1-летних сеянцев сосны обыкновенной. Файл данных (рис. 3) содержит 4 переменных:

VAR1- длина надземной части сеянцев, см;
VAR2- диаметр у корневой шейки, мм;
VAR3- длина корней, см;
VAR4- длина хвои, см;

После выбора процедуры Descriptive statistics на экране появится одноименное диалоговое окно (рис. 4).

Data: PRIMER1.STA 4v * 50c

NUM VAL	Биометрические показатели 1-летних сеянцев сосны с			
	1 VAR1	2 VAR2	3 VAR3	4 VAR4
1	3,6	1,33	20,5	3,100
2	2,3	1,00	19,5	2,500
3	2,9	1,54	14,2	2,600
4	2,6	1,10	14,9	2,500
5	3,1	1,45	14,5	2,600
6	3,2	1,33	15,8	2,500
7	3,5	1,25	11,4	2,800
8	4,2	1,65	17,6	3,300
9	5,0	1,42	19,4	2,600
10	3,0	1,36	21,5	2,700
11	3,1	1,20	17,6	2,800
12	2,9	1,17	18,3	2,700
13	4,0	1,75	17,5	2,300
14	2,2	1,62	21,9	3,500
15	3,1	1,24	16,0	2,400

Рис. 3. Окно файла данных

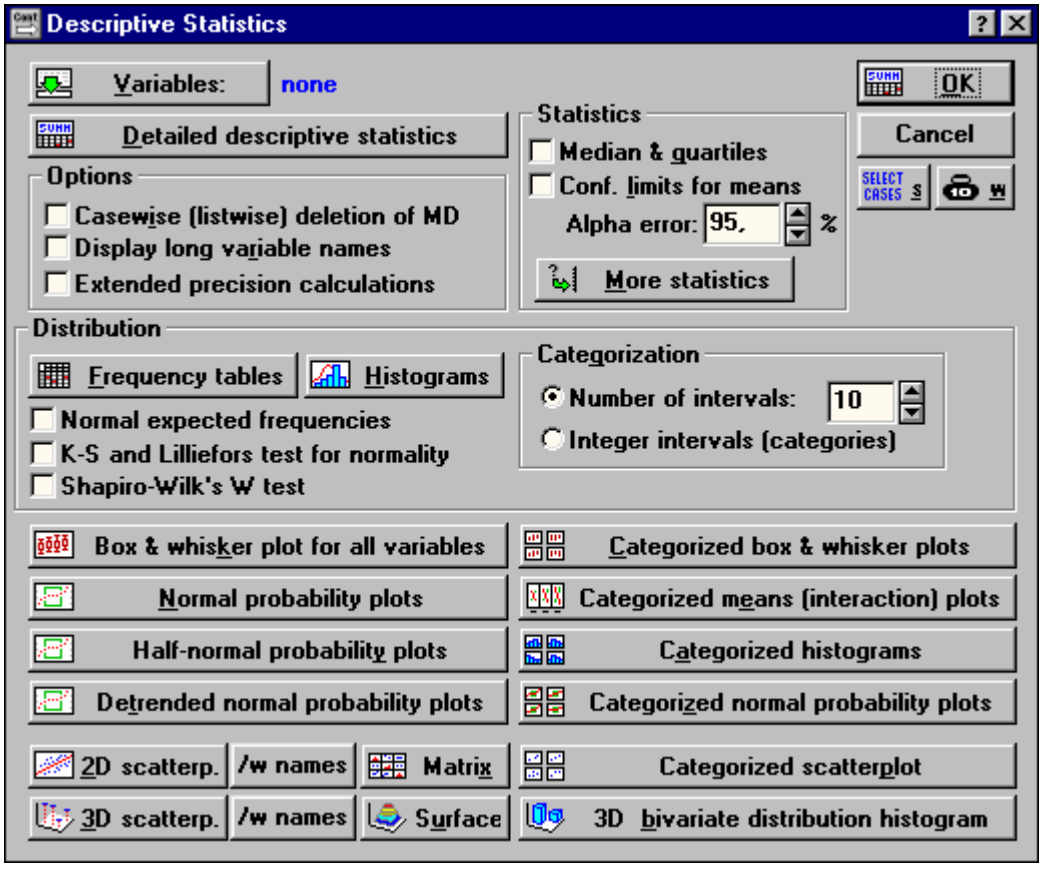


Рис. 4. Диалоговое окно "Descriptive statistics

В этом окне при помощи кнопки **Variables** следует выбрать переменные для анализа (рис.5);

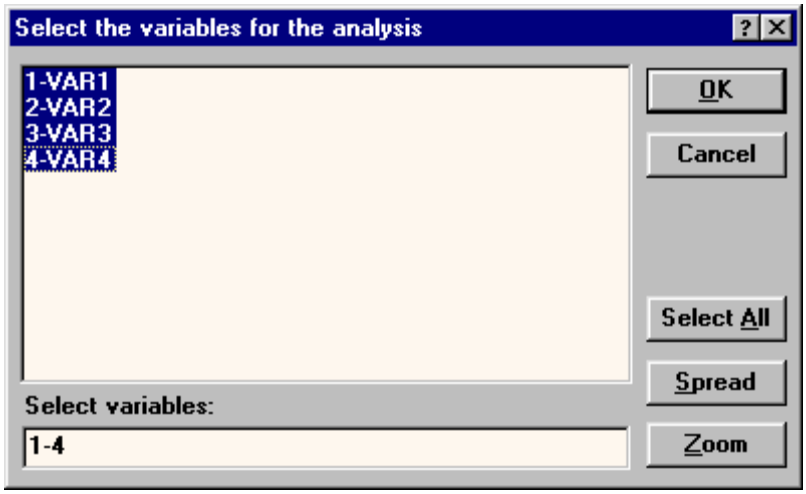


Рис. 5. Окно выбора переменных

На первом этапе обработки данных часто возникает необходимость в их группировке. Группировка позволяет представить первичные данные в компактном виде, выявить закономерности варьирования изучаемого признака. Количество классов можно приблизительно наметить, пользуясь следующими придержками (Лакин, 1990): при количестве наблюдений 25-40 - 5-6 классов, при количестве наблюдений 40-60 - 6-8 классов, 60-100 - 7-10, 100-200 наблюдений - 8-12, более 200 наблюдений - 10-15 классов.

Для построения гистограмм и таблиц частот используется группа кнопок **Distribution** окна Descriptive statistics. Число классов (интервалов) группировки данных устанавливается при помощи счетчика переключателя **Number of intervals** окна Descriptive statistics. Справа от кнопок Distribution находятся две опции **Categorization** (Группировка), позволяющие задать число интервалов группировки или установить величину интервала равную целому числу. Если заактивировать переключатель **Integer intervals (categories)**, то классы (интервалы) группировки будут представлять из себя целые числа.

Результаты группировки длины сеянцев (переменная Var1) представлены в табл. 1.

Таблица 1

Результаты группировки замеров высот

Интервал длин, м	Count (ко-во)	Cumul. Count (ко-во с накоплением)	Percent of Valid (%)	Cumul % of Valid (% с накоплением)	% of all Cases (% от общего ко-ва)
1,0 < x <= 2,0	0	0	0	0	0
2,0 < x <= 3,0	15	15	30	30	30
3,0< x <= 4,0	23	38	46	76	46
4,0 < x <= 5,0	5	43	10	86	10
5,0< x <= 6,0	6	49	12	98	12
6,0 < x <= 7,0	1	50	2	100	2

Представим распределение переменных на гистограммах. Для этого предназаначена кнопка **Histograms** окна Descriptive statistics.

На гистограмму при необходимости можно наложить плотность нормального распределения, проверить близость распределения к нормальному виду при помощи критериев Колмогорова-Смирнова, Лилиефорса; вычислить статистику Шапиро-Уилкса. Для этого в группе опций Distribution необходимо установить флажок напротив соответствующих статистик. Значения статистик показываются прямо на гистограммах.

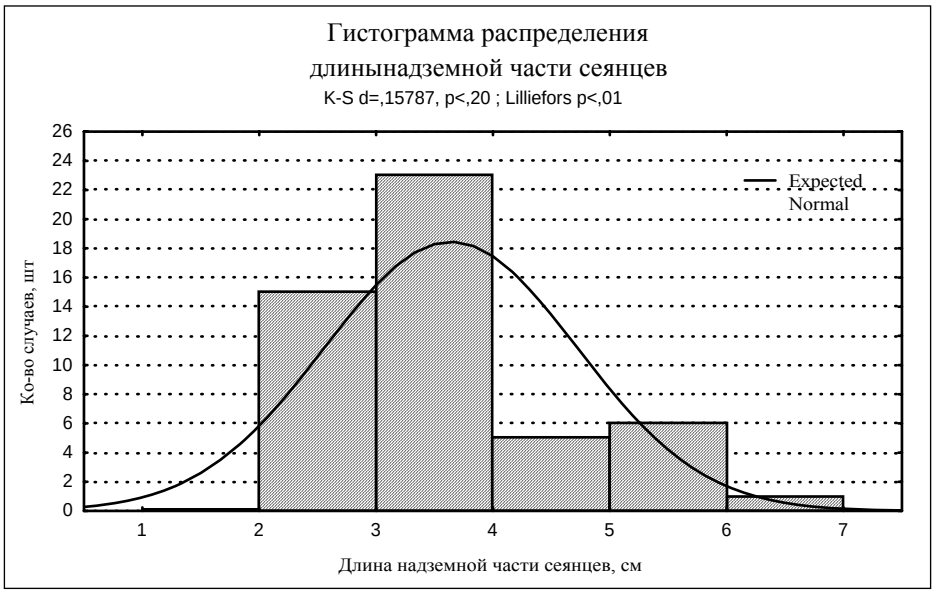


Рис.6. Гистограмма распределения длины надземной части сеянцев

На рис. 6 в качестве примера приводится гистограмма распределения длины надземной части сеянцев (переменной Var1).

На гистограмме показана кривая плотности нормального распределения, а также критерий Колмогорова-Смирнова (d). Статистика Колмогорова-Смирнова оказалась равной 0,157. Чем меньше величина этой статистики, тем ближе распределение случайной величины к нормальному. Вероятность нулевой гипотезы (p) менее 0,20.

О нормальности распределения можно судить по графику на нормальной вероятностой бумаге. Его легко построить при помощи опции **Normal probability plots** окна "Descriptive statistics" (рис.4). Чем ближе распределение к нормальному виду, тем лучше значения ложатся на прямую линию (рис. 7). Этот метод оценки является фактически глазомерным. В сомнительных случаях проверку на нормальность можно продолжить с использованием специальных статистических критериев (Колмогорова-Смирнова, Омега-квадрат (w^2)). Однако детальная проверка гипотезы о нормальности выборки требует довольно значительных объемов выборки (по мнению некоторых авторов не менее 100 наблюдений).

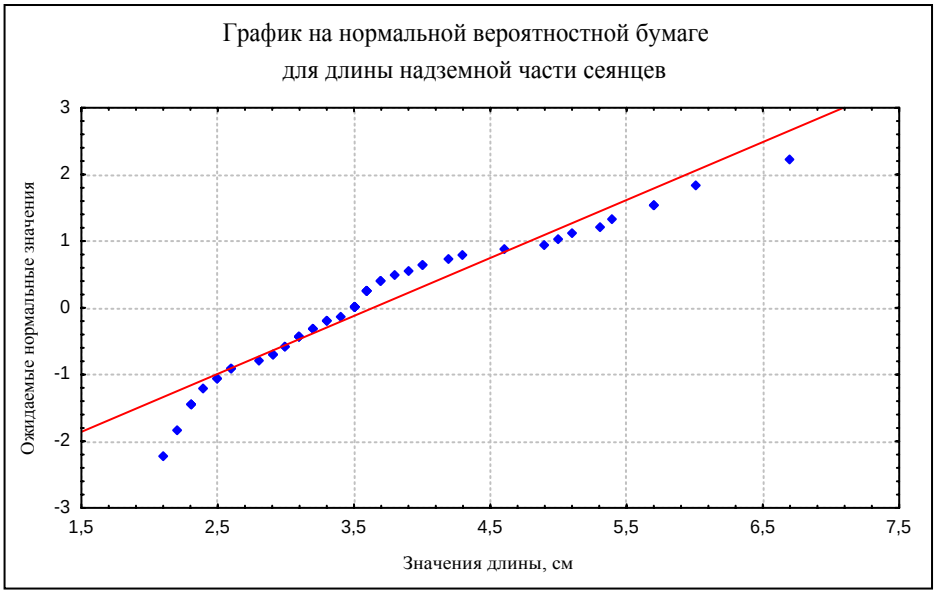


Рис. 7. График на нормальной вероятностной бумаге для выборки длин надземной части сеянцев

Чтобы выбрать статистики, подлежащие вычислению, удобнее всего воспользоваться кнопкой **More statistics** (рис. 8)

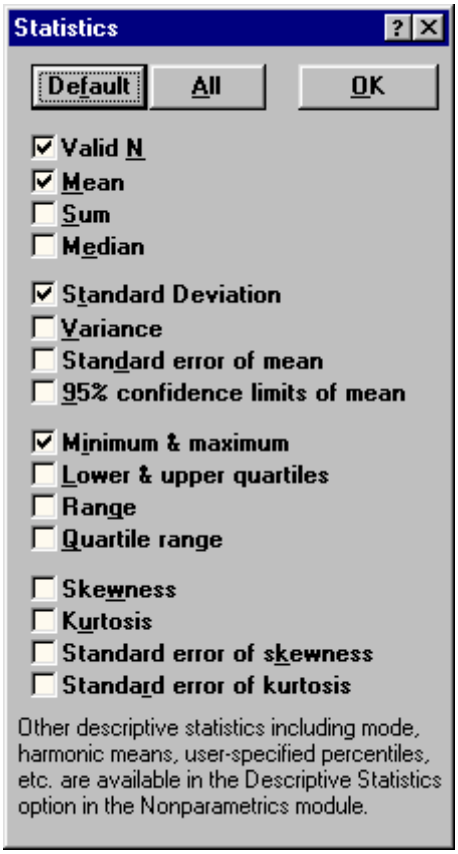


Рис. 8. Окно выбора статистик

Valid N - объем выборки;

Mean - средняя арифметическая;

Среднее значение случайной величины представляет собой наиболее типичное, наиболее вероятное ее значение, своеобразный центр, вокруг которого разбросаны все значения признака.

Sum - сумма;

Median - медиана;

Медианой является такое значение случайной величины, которое разделяет все случаи выборки на две равные по численности части.

Standard Deviation - стандартное отклонение;

Стандартное отклонение (или среднее квадратическое отклонение) является мерой изменчивости (вариации) признака. Оно показывает на какую величину в среднем отклоняются случаи от среднего значения признака. Особенно большое значение имеет при исследовании нормальных распределений. В нормальном распределении 68% всех случаев лежит в интервале $\pm$ одного отклонения от среднего, 95% - $\pm$ двух стандартных отклонений от среднего и 99,7% всех случаев - в интервале $\pm$ трех стандартных отклонений от среднего.

Variance - дисперсия;

Дисперсия является мерой изменчивости, вариации признака и представляет собой средний квадрат отклонений случаев от среднего значения признака. В отличии от

других показателей вариации дисперсия может быть разложена на составные части, что позволяет тем самым оценить влияние различных факторов на вариацию признака. Дисперсия - один из существеннейших показателей, характеризующих явление или процесс, один из основных критериев возможности создания достаточно точных моделей.

Standard error of mean - стандартная ошибка среднего;

Стандартная ошибка среднего это величина, на которую отличается среднее значение выборки от среднего значения генеральной совокупности при условии, что распределение близко к нормальному. С вероятностью 0,68 можно утверждать, что среднее значение генеральной совокупности лежит в интервале $\pm$ одной стандартной ошибки от среднего, с вероятностью 0,95 - в интервале $\pm$ двух стандартных ошибок от среднего и с вероятностью 0,99 - среднее значение генеральной совокупности лежит в интервале $\pm$ трех стандартных ошибок от среднего.

95% confidence limits of mean - 95%-ый доверительный интервал для среднего;

Интервал, в который с вероятностью 0,95 попадает среднее значение признака генеральной совокупности.

Minimum, maximum - минимальное и максимальное значения;

Lower, upper quartiles - нижний и верхний квартили;

Верхний квартиль это такое значение случайной величины, больше которого по величине 25% случаев выборки. Верхний квартиль это такое значение случайной величины, меньше которого по величине 25% случаев выборки.

Range - размах;

Расстояние между наибольшим (maximum) и наименьшим (minimum) значениями признака.

Quartile range - интерквартильная широта;

Расстояние между нижним и верхним квартилями.

Skewness -асимметрия;

Асимметрия характеризует степень смещения вариационного ряда относительно среднего значения по величине и направлению. В симметричной кривой коэффициент асимметрии равен нулю. Если правая ветвь кривой, начиная от вершины) больше левой (правосторонняя асимметрия), то коэффициент асимметрии больше нуля. Если левая ветвь кривой больше правой (левосторонняя асимметрия), то коэффициент асимметрии меньше нуля. Асимметрия менее 0,5 считается малой.

Standard *error of Skewness* -стандартная ошибка асимметрии;

Kurtosis - эксцесс;

Эксцесс характеризует степень концентрации случаев вокруг среднего значения и является своеобразной мерой крутости кривой. В кривой нормального распределения эксцесс равен нулю. Если эксцесс больше нуля, то кривая распределения характеризуется островершинностью, т.е. является более крутой по сравнению с нормальной, а случаи более густо группируются вокруг среднего. При отрицательном эксцессе кривая является более плосковершинной, т.е. более пологой по сравнению с нормальным распределением. Отрицательным пределом величины эксцесса является число -2, положительного предела - нет.

Standard error of Kurtosis - стандартная ошибка эксцесса.

На против статистик, подлежащих вычислению (рис. 8) следует поставить флажок.

После нажатия на кнопку OK окна Descriptive statistics на экране появится таблица с результатами расчетов описательных статистик (рис. 9).

Descriptive Statistics (primer1.sta)

Continue...	Valid N	Mean	Confid. -95,000	Confid. +95,000	Median	Minimum	Maximum	Lower Quartil
VAR1	50	3,64	3,33	3,95	3,50	2,100	6,70	2,90
VAR2	50	1,15	1,06	1,24	1,15	,500	1,76	,96
VAR3	50	16,97	15,67	18,27	17,70	4,700	26,50	15,70
VAR4	50	2,55	2,42	2,67	2,50	1,600	3,60	2,20

Рис. 9. Окно с результатами расчета описательных статистик

В таблице 2 эти данные представлены после копирования в текстовый редактор Word.

К сожалению, пакет Statistica не рассчитывает такие часто применяемые статистики, как коэффициент вариации и относительная ошибка среднего значения (точность опыта). Но их определение не представляет большого труда. Коэффициент вариации (%) есть отношение стандартного отклонения к среднему значению, умноженное на 100%:

$$\text{КоэффициентВариации} = \frac{\text{StandardDeviation}}{\text{Mean}} \cdot 100\%$$

Коэффициент вариации, как дисперсия и стандартное отклонение, является показателем изменчивости признака. Коэффициент вариации не зависит от единиц измерения, поэтому удобен для сравнительной оценки различных статистических совокупностей. При величине коэффициента вариации до 10% изменчивость оценивается как слабая, 11-25% - средняя, более 25% - сильная (Лакин, 1990).

Относительная ошибка среднего значения (%) - отношение стандартной ошибки среднего к среднему значению, умноженное на 100% (для вероятности 0,68):

$$\text{ОтносительнаяОшибкаСреднегоЗначения} = \frac{\text{StandardErrorOfMean}}{\text{Mean}} \cdot 100\%$$

Это процент расхождения между генеральной и выборочной средней, показывает на сколько процентов можно ошибиться, если утверждать, что генеральная средняя равна выборочной средней. Если относительная ошибка не превышает 5%, то точность исследований (точность опыта) оценивается как хорошая, до 10% - удовлетворительная.

Точность 3-5% при вероятности 0,95, а в некоторых случаях и при вероятности 0,68, является вполне достаточной для большинства задач лесного хозяйства.

Таблица 2

Основные описательные статистики выборки 1-летних сеянцев сосны обыкновенной

Переменная	Valid N	Mean	Confid. -95%	Confid. +95%	Median	Minimum	Maximum	Lower Quartile	Upper Quartile
VAR1	50	3,64	3,33	3,95	3,50	2,1	6,70	2,90	4,00
VAR2	50	1,15	1,06	1,24	1,15	0,5	1,76	0,96	1,37
VAR3	50	16,97	15,67	18,27	17,70	4,7	26,50	15,70	19,70
VAR4	50	2,55	2,42	2,67	2,50	1,6	3,60	2,20	2,80

Переменная	Range	Quartile Range	Variance	Std.Dev.	Standard Error	Skewness	Std.Err. Skewness	Kurtosis	Std.Err. Kurtosis
VAR1	4,60	1,10	1,169	1,081	0,153	0,921	0,337	0,403	0,662
VAR2	1,26	0,41	0,098	0,313	0,044	-0,080	0,337	-0,451	0,662
VAR3	21,80	4,00	20,865	4,568	0,646	-0,834	0,337	0,772	0,662
VAR4	2,00	0,60	0,200	0,447	0,063	0,386	0,337	0,036	0,662

При необходимости обработки сгруппированных данных нужно воспользоваться кнопкой **Weight** окна Descriptive statistics (рис.4). В появляющемся диалоговом окне (рис. 10) следует указать переменную, являющуюся ве-

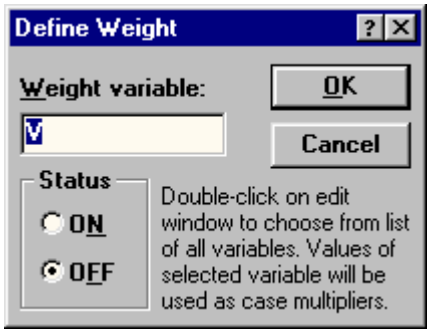


Рис.106. Окно задания переменной- весов

сами для других переменных (Weight variables), а переключатель Status установить в положение ON. Необходимо иметь в виду, что весы действуют сразу для всех переменных. Поэтому обрабатывать сгруппированные и не сгруппированные данные нужно отдельно.

При помощи опции **Alpha error** (рис. 4) выбирается уровень доверительной вероятности статистического анализа. В биологических исследованиях наиболее часто используется вероятность 0,95 (95%). Вероятности 0,95 соответствует уровень значимости 0,05 (5%).

Кнопка **Select cases** позволяет установить условия включения (include if) или исключения (exclude if) случаев (строк файла данных) из статистической обработки (рис. 11). Операторы, которые могут использоваться при написании выражений, а также примеры самих выражений имеются непосредственно на самом диалоговом окне Case Selection Conditions (рис. 11) в нижней его части.

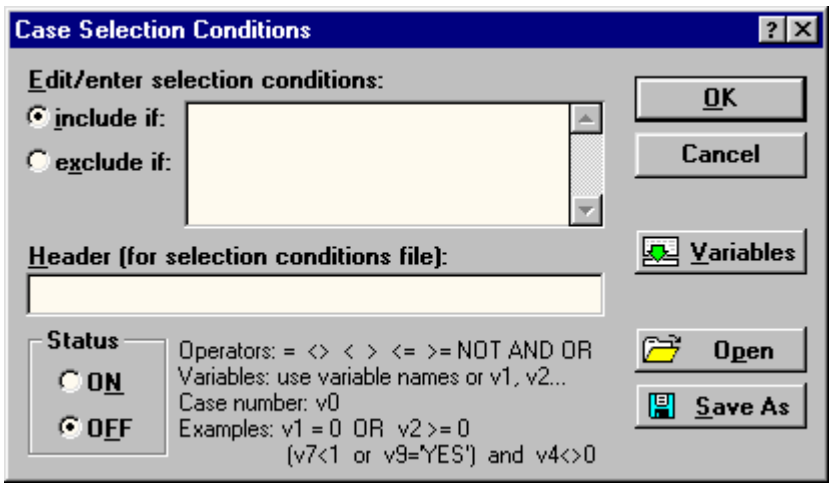


Рис. 11. Окно задания условий выбора случаев

Для визуализации описательных статистик можно построить статистические графики типа "коробок" (или "ящиков с усами"). Это легко можно сделать при помощи кнопки **Box & Whisker plot for all variable** окна Descriptive statistics. На графике можно отобразить 3 статистики, установив переключатель в одно из 4-х положений (рис. 12):

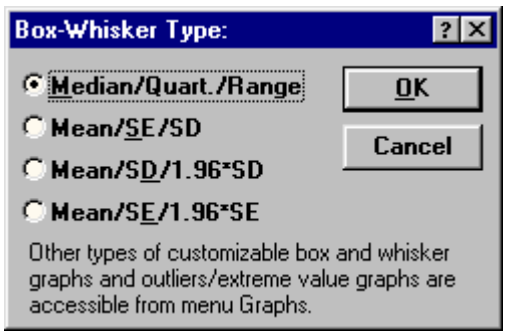


Рис. 12. Окно выбора статистик для графика коробок

1. ***Median/Quart./Range*** - Медиана / Квартили / Размах;
2. ***Mean/SE/SD*** - Среднее / Ошибка среднего / Стандартное отклонение
3. ***Mean/SD/1.96SD*** - Среднее / Стандартное отклонение / Интервал 1,96* стандартного отклонения;
4. ***Mean/SE/1.96*SE*** - Среднее / Ошибка среднего / Интервал 1,96 * ошибки среднего.

Визуализация описательных статистик переменных VAR1, VAR3 и VAR4 рассматриваемого примера при помощи графика коробок представлена на рис. 13.

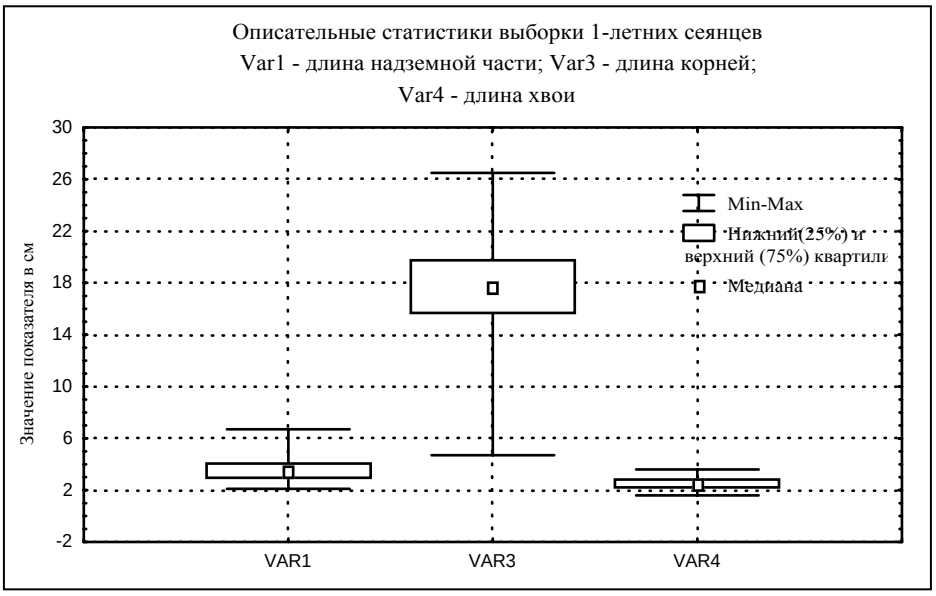


Рис. 13. Описательные статистики в графическом виде

2.2. Процедура Correlation matrices (Корреляционные матрицы)

Эта процедура предназначена для проведения корреляционного анализа, установления тесноты линейной связи между переменными.

Установим тесноту взаимосвязей между таксационными показателям дубовых древостоев. Фрагмент окна файла данных представлен на рис. 14. Данные представляют собой таксационные показатели древостоев 93 пробных

площадей, заложенных в низкоствольных дубравах 4 класса бонитета. По названию переменных понятно какие таксационные показатели они содержат.

Data: DUB4.STA 16v * 93c

Таксационные показатели древостояев порослевого дуба 4 кла

	2 %ДУБА	3 N_ДУБА	4 G_ДУБА	5 D_ДУБА	6 H_ДУБА	7 A	8 M_ДУБА
1	50	1710	4,47	5,7	4,8	14	12,68
2	98	813	18,68	17,1	13,7	56	133,00
3	88	797	21,46	18,3	15,4	68	149,30
4	97	1258	26,51	16,4	14,1	56	190,13
5	36	311	7,25	17,2	14,8	57	52,80
6	85	1062	18,90	15,1	14,0	47	127,30
7	95	1150	21,26	15,3	11,2	38	131,40
8	92	873	20,56	17,3	13,8	55	139,40
9	61	1620	3,48	5,2	5,7	16	15,00
10	56	1790	5,27	6,1	5,2	14	16,10
11	54	1210	7,80	9,0	7,5	19	26,60
12	69	300	9,06	19,6	14,4	67	65,65
13	100	1145	21,72	15,6	12,8	54	154,50
14	48	278	10,85	22,3	15,4	61	90,00
15	77	1105	12,56	12,0	10,0	38	52,70

Рис.14. Окно файла данных

В стартовом окне этой процедуры **"Pearson Product-Moment Correlation"** (Корреляция Пирсона) (рис. 15) для расчета квадратной матрицы используется кнопка **One variable list (square matrix).**

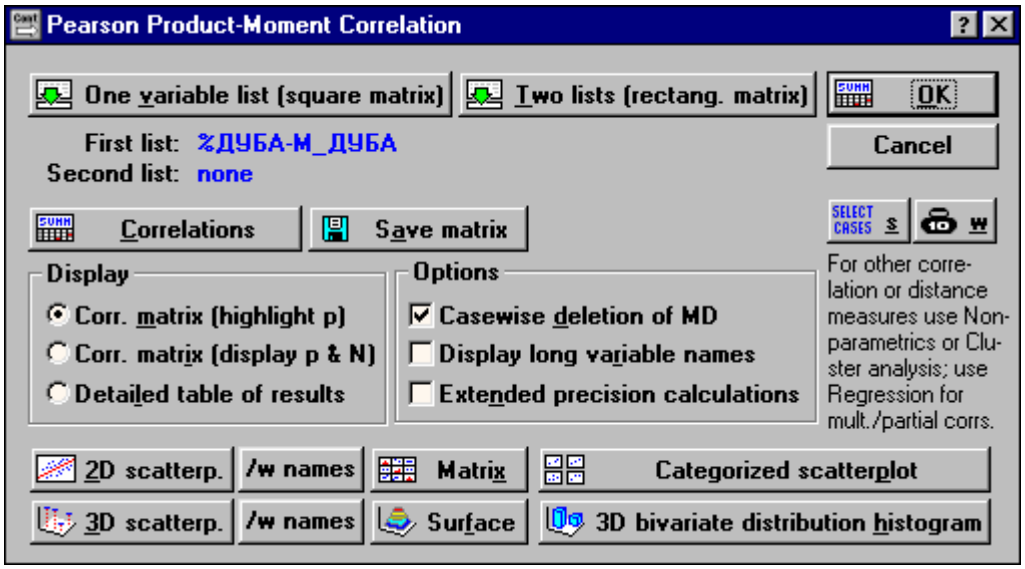


Рис. 15. Окно Pearson Product-Moment Correlation

В списке переменных выбирают переменные, между которыми будут рассчитаны парные коэффициенты корреляции Пирсона. После нажатия на кнопку OK или Correlationes на экране появится корреляционная матрица (рис. 16).

Correlations (dub4.sta)

Continue...

Marked correlations are significant at p < ,05000
N=93 (Casewise deletion of missing data)

Variable	%ДУБА	N_ДУБА	G_ДУБА	D_ДУБА	H_ДУБА	A	M_ДУБА
%ДУБА	1,00	,49	,63	-,03	,02	,02	,50
N_ДУБА	,49	1,00	,18	-,74	-,68	-,68	-,10
G_ДУБА	,63	,18	1,00	,38	,46	,45	,94
D_ДУБА	-,03	-,74	,38	1,00	,95	,95	,59
H_ДУБА	,02	-,68	,46	,95	1,00	,95	,67
A	,02	-,68	,45	,95	,95	1,00	,67
M_ДУБА	,50	-,10	,94	,59	,67	,67	1,00

Рис. 16. Корреляционная матрица

Коэффициент корреляции - это показатель, оценивающий тесноту линейной связи между признаками. Он может принимать значения от -1 до +1. Знак "-" означает, что связь обратная, "+" - прямая. Чем ближе коэффициент к $|1|$, тем теснее линейная связь. При величине коэффициента корреляции (по Дворецкому) менее 0,3 связь оценивается как слабая, от 0,31 до 0,5 - умеренная, от 0,51 до 0,7 - значительная, от 0,71 до 0,9 - тесная, 0,91 и выше - очень тесная. Для практических целей Дворецкий рекомендует использовать значительные, тесные и очень тесные связи.

Процедура Correlation matrices сразу же дает возможность проверить достоверность рассчитанных коэффициентов корреляции. Значение коэффициента корреляции может быть высоким, но не достоверным, случайным. Чтобы увидеть вероятность нулевой гипотезы (p), гласящей о том что коэффициент корреляции равен 0, нужно в опции **Display** окна Pearson Product-Moment Correlation (рис. 15) установить переключатель на вторую строку **Corr. matrix (display p & N)**. Но даже если этого не делать и оставить переключатель в первом положении **Corr. matrix (highlight p),** статистически значимые на 5-% уровне коэффициенты корреляции будут выделены в корреляционной матрице на экране монитора цветом, а при распечатке помечены звездочкой. Третье положение переключателя опции Display **- Detail table of results** позволяет просмотреть результаты корреляционного анализа в деталях (рис. 17). Флажок опции **Casewise deletion of MD** устанавливается для исключения из обработки всей строки файла данных, в которой есть хотя бы одно пропущенное значение.

Correlations (dub4.sta)

Continue...

Marked correlations are significant at p < ,05000
(Casewise deletion of missing data)

Var. X & Var. Y	Mean	Std.Dv.	r(X,Y)	rI	t	p	N
N_ДУБА	960,60	514,26					
G_ДУБА	13,81	5,74	,1777	,0316	1,72	,0884	93
N_ДУБА	960,60	514,26					
D_ДУБА	14,65	4,33	-,7396	,5469	-10,48	,0000	93
N_ДУБА	960,60	514,26					
H_ДУБА	12,06	2,94	-,6804	,4630	-8,86	,0000	93
N_ДУБА	960,60	514,26					
A	45,69	15,50	-,6770	,4583	-8,78	,0000	93

Рис. 17. Вариант детального просмотра результатов корреляционного анализа

2.3. Процедура t-test for independent samples (t-критерий для независимых выборок)

Эта процедура используется для установления достоверной статистической разницы между средними значениями выборок на основе t-критерия Стьюдента.

Имеются результаты определения водопроницаемости почвы на площадках с различным характером напочвенного покрова (табл. 3). Создадим файл с данными с четырьмя переменными:

VAR1 - Водопроницаемость на площадке 1 (Мертвый покров, лесная подстилка 2.5см)

VAR2 - Водопроницаемость на площадке 2 (Травяной покров, проективное покрытие 40-50%, задернение 10%)

VAR3 - Водопроницаемость на площадке 3 (Травяной покров, проективное покрытие 100%, задернение 70%)

VAR4 - Водопроницаемость на площадке 4 (Травяной покров, проективное покрытие 30-40%, задернения нет)

Таблица 3

Значения переменных VAR1, VAR2, VAR3, VAR4
(Водопроницаемость почвы (мм/мин) в зависимости от характера напочвенного покрова)

Переменная			
VAR1	VAR2	VAR3	VAR4
1	2	3	4
303	78,7	53,5	67,9
238	82	68	105,3
303	58,1	38,8	149,3
238	97,1	49,5	138,9
303	73	70,4	45,5

Продолжение табл. 3

1	2	3	4
200	142,9	40,5	98
400	55,6	25,1	61,3
238	108,7	12,2	75,8
263	69,9	33,6	71,4
303	120,5	28,3	35,7

Окно с файлом данных этого примера приводится на рис. 18.

Data: VODA.STA 4v * 10c

NUM VAL	Водопроницаемость почв (мм/ми) в зависимости от х			
	1 VAR1	2 VAR2	3 VAR3	4 VAR4
1	303,0	78,7	53,5	67,9
2	238,0	82,0	68,0	105,3
3	303,0	58,1	38,8	149,3
4	238,0	97,1	49,5	138,9
5	303,0	73,0	70,4	45,5
6	200,0	142,9	40,5	98,0
7	400,0	55,6	25,1	61,3
8	238,0	108,7	12,2	75,8
9	263,0	69,9	33,6	71,4
10	303,0	120,5	28,3	35,7

Рис.18. Окно с файлом данных

Влияет ли характер напочвенного покрова на водопроницаемость почвы с ее поверхности? Воспользуемся процедурой t-test for independent samples для расчета средних величин водопроницаемости по вариантам опыта и одновременно проверим достоверность различий между средними значениями.

В окне "T-Test for independent samples (Groups)" (рис. 19) в опции **Input file** следует указать тип файла с данными:

- **One record percase (use a grouping variable)** - одна запись на случай (используя группирующую переменную);
- **Each variable contains the data for one group** - каждая переменная содержит данные одной группы.

Использующийся нами файл данных (рис.) относится ко второму типу (Each variable contains the data for one group).

При помощи кнопки **Variables** выбираются переменные для по парного сравнения. При этом должны быть выбраны переменные в обоих списках. Чтобы сравнить попарно сразу все варианты опыта друг с другом, следует выбрать переменные так, как показано на рис. 20.

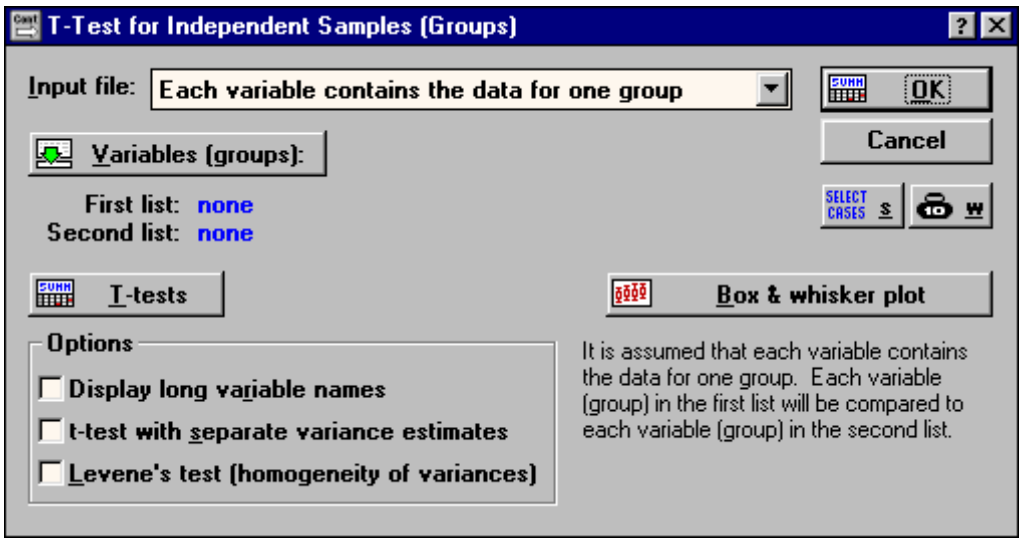


Рис. 19. Окно "T-Test for independent samples (Groups)"

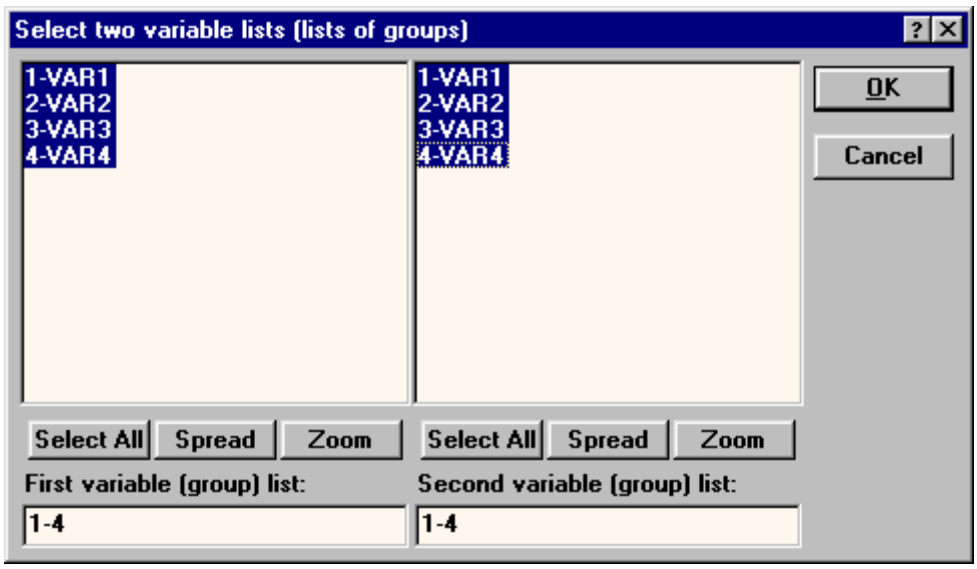
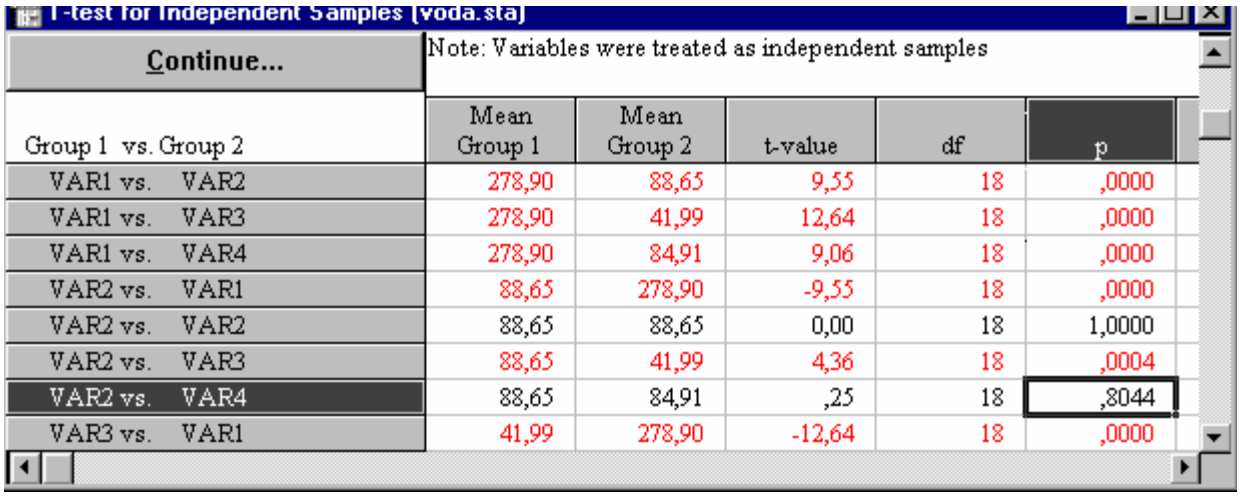


Рис.20. Выбор переменных для по парного сравнения

После нажатия на кнопку **OK** или **T-test** на экране появляется таблица с результатами сравнения по t-критерию.

Фрагмент окна с результатами проведения процедуры приводится на рис. 21. Согласно нулевой гипотезы между средними значениями водопроницаемости достоверного различия нет, т.е. две выборки однородны и представляют одну генеральную совокупность. Если вероятность нулевой гипотезы (p) меньше 5% (т.е. $p < 0,05$), то с вероятность 0,95 нулевую гипотезу можно отбросить. По парное сравнение средних величин водопроницаемости показало достоверное различие между всеми вариантами опыта, кроме вариантов 2 и 4. Нулевую гипотезу в последнем случае отбросить нельзя, так как ее вероятность чересчур высока (p=0,804).


T-test for Independent Samples (voda.sta)

Continue...

Note: Variables were treated as independent samples

Group 1 vs. Group 2	Mean Group 1	Mean Group 2	t-value	df	p
VAR1 vs. VAR2	278,90	88,65	9,55	18	,0000
VAR1 vs. VAR3	278,90	41,99	12,64	18	,0000
VAR1 vs. VAR4	278,90	84,91	9,06	18	,0000
VAR2 vs. VAR1	88,65	278,90	-9,55	18	,0000
VAR2 vs. VAR2	88,65	88,65	0,00	18	1,0000
VAR2 vs. VAR3	88,65	41,99	4,36	18	,0004
VAR2 vs. VAR4	88,65	84,91	,25	18	,8044
VAR3 vs. VAR1	41,99	278,90	-12,64	18	,0000

Рис.21. Результаты проведения процедуры t-test for independent samples

2.4. Процедура Breakdown / one way ANOVA (Классификация и однофакторный дисперсионный анализ)

Эта процедура используется для проведения простейшего варианта однофакторного дисперсионного анализа данных по схеме полной рендомизации (неорганизованных повторений). Не позволяя вычленить дисперсию блоков (повторений), рядов, столбцов, процедура не предназначена для обработки данных, полученных по активным опытным схемам (рендомизированных блоков, смехе латинского квадрата, расщепленных делянок и блоков).

Воспользуемся исходными данными примера из раздела 2.3. и, проведя дисперсионный анализ, выясним влияет ли характер напочвенного покрова на водопроницаемость почв с ее поверхности. Для проведения процедуры Breakdown / one way ANOVA следует создать файл с данными из двух переменных (табл.):

VAR1 - Водопроницаемость почвы с поверхности (мм/мин) по всем вариантам опыта

VAR2 - Номер варианта опыта (1, 2, 3 или 4)

Таблица 4

Значения переменных VAR1 и VAR2

VAR1	VAR2	VAR1	VAR2	VAR1	VAR2	VAR1	VAR2
303	1	78,7	2	53,5	3	67,9	4
238	1	82	2	68	3	105,3	4
303	1	58,1	2	38,8	3	149,3	4
238	1	97,1	2	49,5	3	138,9	4
303	1	73	2	70,4	3	45,5	4
200	1	142,9	2	40,5	3	98	4
400	1	55,6	2	25,1	3	61,3	4
238	1	108,7	2	12,2	3	75,8	4
263	1	69,9	2	33,6	3	71,4	4
303	1	120,5	2	28,3	3	35,7	4

На рис. 22 представлен вид окна с файлом данных.

В окне Descriptive Statistics and Correlations by Groups (Breakdown) (рис. 23) в опции **Analysis** следует выбрать: **Detailed analysis of individual tables**. Вторая строка в списке Analysis - Bach process (and print) list of table предназначена для создания таблицы частот сгруппированных данных и разбитых на интервалы зависимых переменных. Флажок опции **Casewise (listwise) deletion of MD** устанавливается для исключения из обработки всей строки файла данных, в которой есть хотя бы одно пропущенное значение.

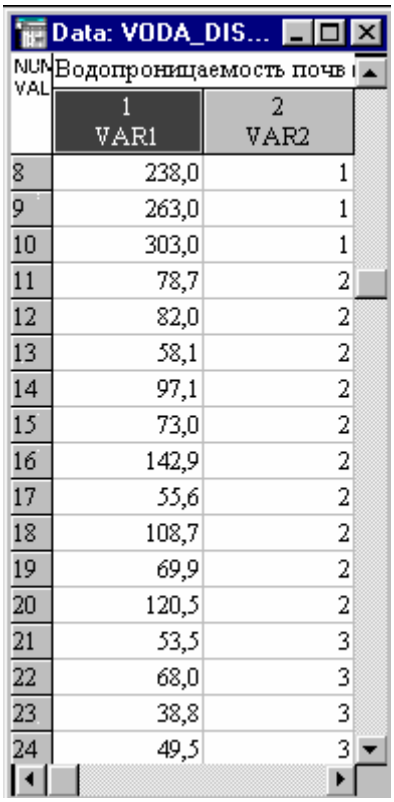


Рис. 22. Вид окна файла данных

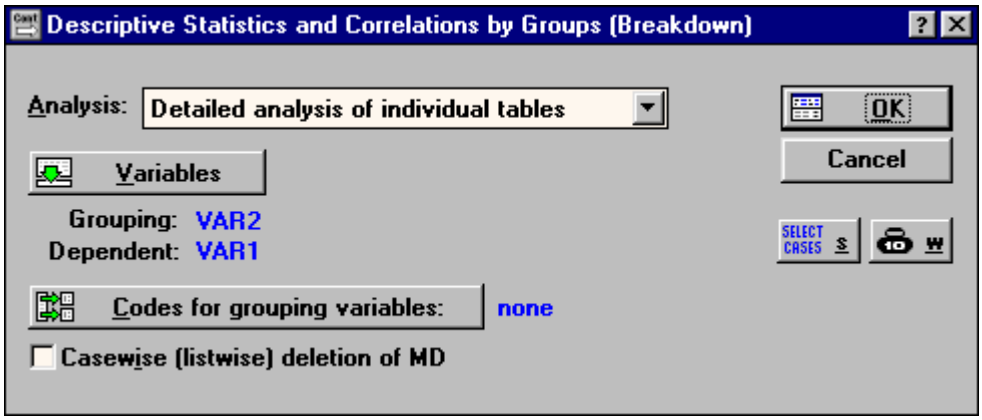


Рис. 23. Окно "Descriptive Statistics and Correlations by Groups (Breakdown)"

Через кнопку **Variables** выбирается зависимая переменная (Dependent variables) и группирующая переменная (Grouping variables), с помощью которой случаи будут разбиты на группы. Группирующей (Grouping) переменной в нашем примере является переменная VAR2, с ее помощью данные по водопроницаемости из зависимой (Dependent) переменной VAR1 группируются по четырем вариантам опыта.

После возвращения в диалоговое окно Descriptive Statistics and Correlations by Groups (Breakdown) и нажатия на кнопку **ОК** на экране появится окно Results (рис. 24) с результатами дисперсионного анализа. При помощи кнопок и опций этого окна в удобном виде можно просмотреть результаты обработки сгруппированных данных.

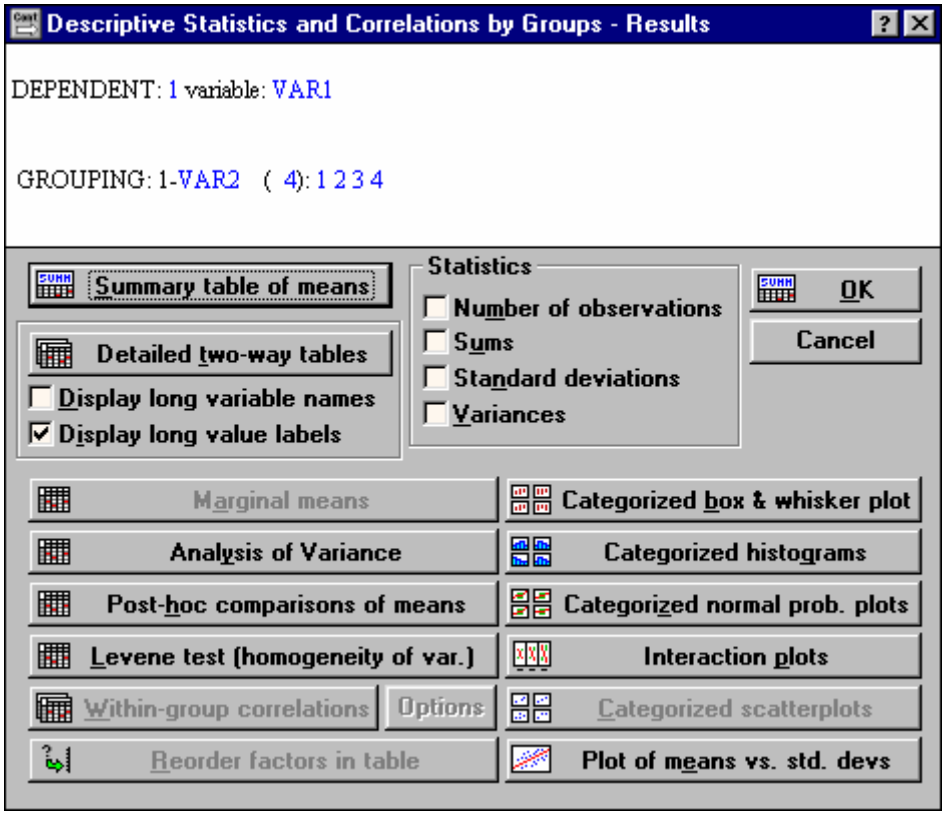


Рис.24 . Окно Results процедуры Breakdown

Дисперсионный анализ заключается в разложении общей изменчивости признака на составные части: с одной стороны на вариацию, определяемую действием изучаемого конкретного фактора, а с другой стороны - вариацию, вызываемую случайными, неконтролируемыми в данном опыте факторами. Основные результаты дисперсионного анализа и проверку нулевой гипотезы однофакторного дисперсионного анализа (утверждающей, что фактор не влияет на вариацию зависимой переменной, т.е. вся вариация сводится к случайной) можно просмотреть при помощи кнопки **Analysis of Variance** (Анализ дисперсий) окна Result (табл. 5).

Таблица 5

Результаты дисперсионного анализа

	SS Effect (сумма квадратов фактора)	df Effect (число степеней свободы фактора)	MS Effect (средний квадрат фактора)	SS Error (сумма квадратов ошибки)	df Error (число степеней свободы ошибки)	MS Error (средний квадрат ошибки)	F	p (вероятность нулевой гипотезы)
VAR1	334967	3	111655,67	51531,70	36	1431,436	78,00	0,0000000

Проверка нулевой гипотезы осуществляется при помощи F-критерия (Критерия Фишера). F- критерий используется как общий критерий, подтверждающий или опровергающий значимое влияние фактора на общую вариацию признака. В нашем примере низкая вероятность нулевой гипотезы (p=0,000000) позволяет ее отвергнуть и говорить о достоверном влиянии характера напочвенного покрова на водопроницаемость почвы.

Просмотрим средние значения водопроницаемости по вариантам опыта при помощи кнопки **Summary tables of means** окна Results (табл. 6).

Таблица 6

Средние значения водопроницаемости по вариантам опыта

Вариант опыта	Водопроницаемость, мм/мин
1	278,90
2	88,65
3	41,99
4	84,91

Не смотря на то, что значимое влияние фактора доказано, это автоматически не означает, что каждый вариант опыта существенно отличается от всех других. Поэтому следующим важным этапом дисперсионного анализа является установление существенности частных различий, т.е. сравнение средних значений водопроницаемости по вариантам опыта. Для этого используется процедура **Post-hot comparisons of means** (Post-hot сравнения средних) (рис. 25). Сравнение групповых средних может производиться при помощи различных критериев:

LSD test of planned comparisons - LSD - тест плановых сравнений. Этот критерий сравнения в отечественной литературе по статистике известен как наименьшая существенная разница (НСР).

Scheffii test - тест Шеффе.

Tukey (HSD) test - тест Тьюки. Тесты Шеффе и Тьюки считаются устаревшими (Литтл, Хиллз, 1981).

Duncan's multiple range test & critical ranges - Многоранговый критерий Дункана.

Выбор критериев осуществляется в диалоговом окне Post-hot Comparisons of Means (рис. 25)

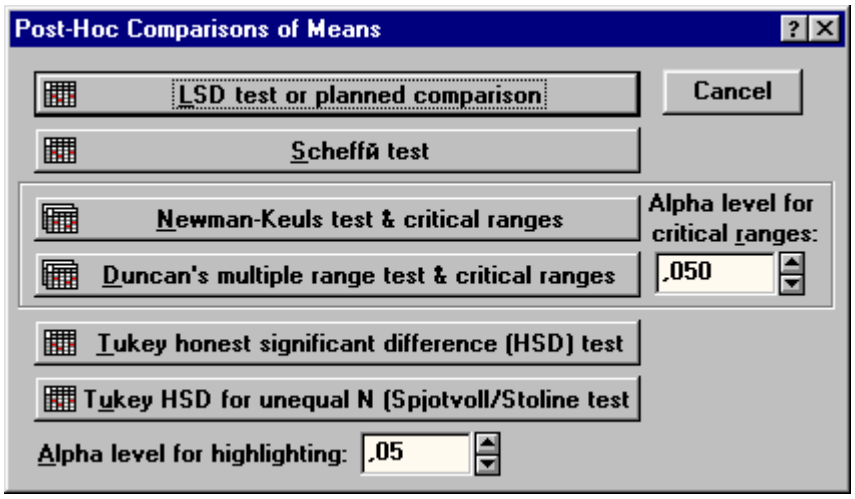


Рис.25 . Диалоговое окно Post-hot Comparisons of Means

Проведем сравнение средних значений водопроницаемости по вариантам опыта при помощи такого широко применяемого точечного критерия как HCP (LSD test of planned comparisons) (рис. 26).

Анализируя результаты теста, представляющие собой вероятность нулевой гипотезы по парного сравнения средних величин водопроницаемости, мы видим достоверное различие на 5%-ом уровне между всеми вариантами опыта, кроме вариантов 2 и 4. Нулевую гипотезу в последнем случае отбросить нельзя, так как ее вероятность высока (p=0,826).

LSD Test; Variable: VAR1 (voda_dis.sta)

Continue... | Marked differences are significant at p < ,05000

VAR2	{1} M=278,90	{2} M=88,650	{3} M=41,990	{4} M=84,910
G_1:1 {1}		,000000	,000000	,000000
G_2:2 {2}	,000000		,009087	,826310
G_3:3 {3}	,000000	,009087		,015671
G_4:4 {4}	,000000	,826310	,015671	

Рис.26. Результаты сравнения групповых средних по HCP

Сама величина HCP на экран не выводится, но если она потребуется, то она может быть легко рассчитана:

$$HCP_{0,05} = t_{0,05}\sqrt{\frac{2 \cdot MSError}{n}}$$

где: $t_{0,05}$ - величина t - критерия для 5%-ного уровня значимости (определяемся для числа степеней свободы, равному df Error); MS Error - средний квадрат ошибки; n - повторность опыта.

В нашем примере: $HCP_{0,05} = 2,04\sqrt{\frac{2 \cdot 1431,36}{10}} = 34,5$

ПРОВЕДЕНИЕ РЕГРЕССИОННОГО АНАЛИЗА ПРИ ПОМОЩИ МОДУЛЯ Multiple Regressions

В стартовом диалоговом окне этого модуля (рис. 27.) при помощи кнопки **Variables** указываются зависимая (dependent) и независимые (ая) (independent) переменные. В поле **Input file** указывается тип файла с данными:

Raw Date - данные в виде строчной таблицы;

Correlation Matrix - данные в виде корреляционной матрицы.

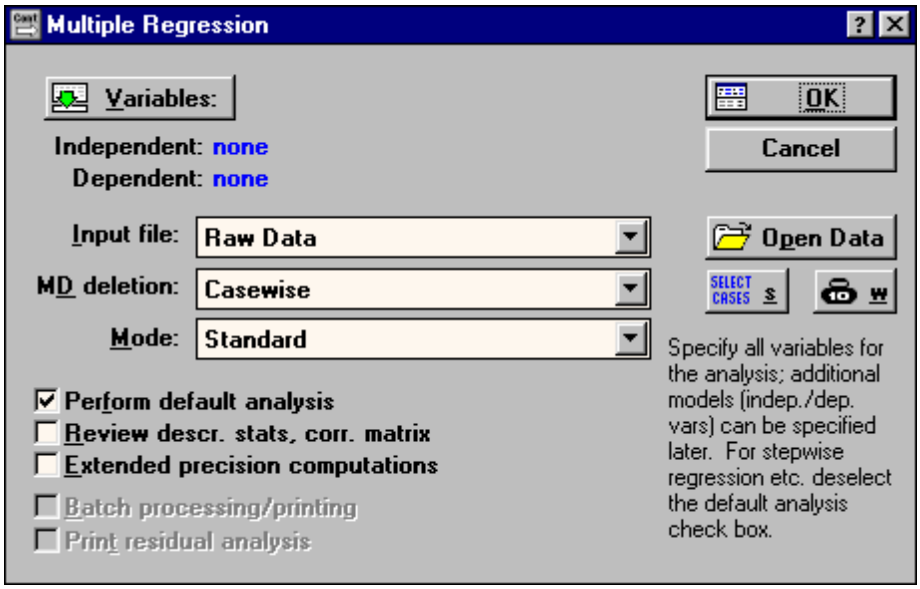


Рис.27 . Стартовое диалоговое окно модуля Multiple Regressions

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

casewise - игнорируется вся строка, в которой есть хотя бы одной пропущенное значение;

mean Substitution - взамен пропущенных данных подставляются средние значения переменных;

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

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

Standard - стандартная линейная модель вида:

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

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

Рассмотрим проведение регрессионного анализа на примере. Имеются данные обмера и таксации 380 модельных деревьев различных древесных пород. В файле данных (рис. 30) 10 переменных:

1	PORODA	Древесня порода (d- дуб, lp- липа, k- клен, o - осина)
2	A	Возраст дерева, лет

3	D	Таксационный диаметр ствола дерева в коре, см
4	H	Высота дерева, м
5	VK	Объем ствола в коре, куб.м
6	V	Объем ствола без коры, куб.м
7	Q2	Второй коэффициент формы
8	L	Длина кроны дерева, м
9	DKR	Диаметр кроны дерева, м
10	F	Старое видовое число

Data: MODL.STA 10v * 380c

TEXT VALUE: Таксационные показатели модельных деревьев

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

Рис.30. Вид окна с файлом данных

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

Выставим опции стартового окна регрессионного анализа (рис.29):

Variables: зависимая (dependent) переменная - VK; независимые (independent) - D,H (рис. 31); **Input file - Raw Date** (данные файла в виде строчной таблицы); **MD deletion - pairwise; Mode - Standard.**

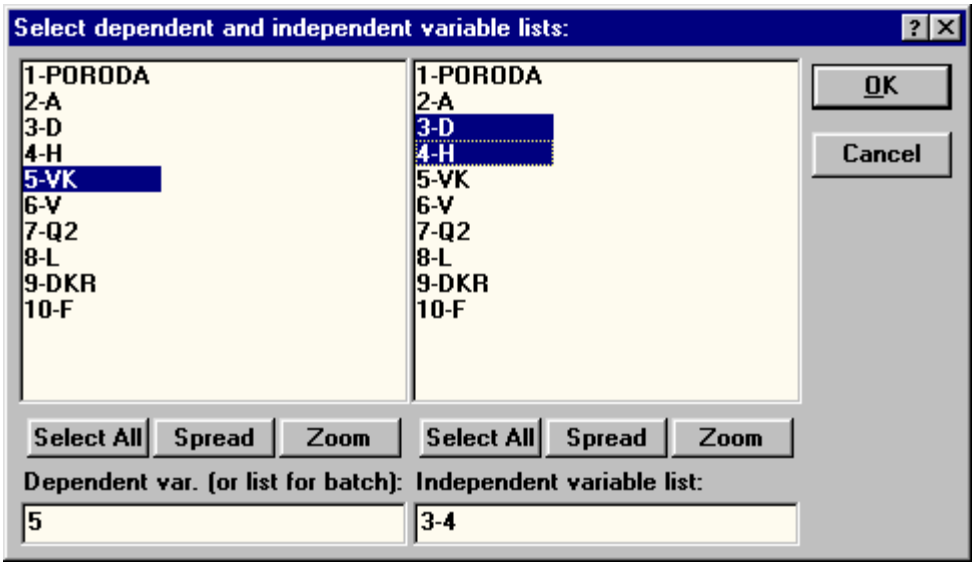


Рис. 31. Выбор зависимой и независимых переменных

Так как в файле данных содержится информация о модельных деревьях разных пород, а уравнение регрессии мы хотим получить для дуба, нужно воспользоваться кнопкой Select cases диалогового окна Multiple Regressions чтобы установить условие включения случаев (строк файла данных) в статистическую обработку. В обработку должны включаться только те строки файла данных, для которых значение первой переменной V1 = 'd' (т.е. дуб) (рис. 32).

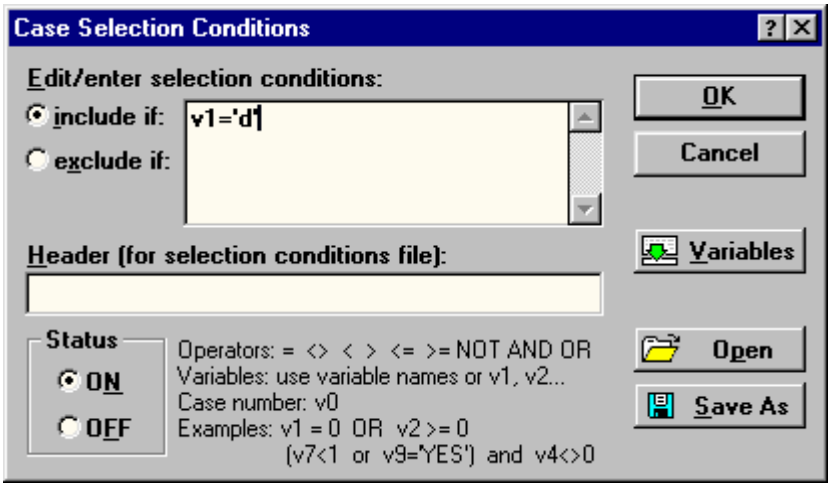


Рис. 32. Задание условия включения в обработку случаев со значением переменной V1 - дуб

После того, как все опции стартового диалогового окна регрессионного анализа выставлены, нажатие на кнопку OK приведет к появлению окна Multiple Regressions Results (результаты регрессионного анализа) (рис. 33), с помощью которого можно просмотреть результаты анализа в деталях.

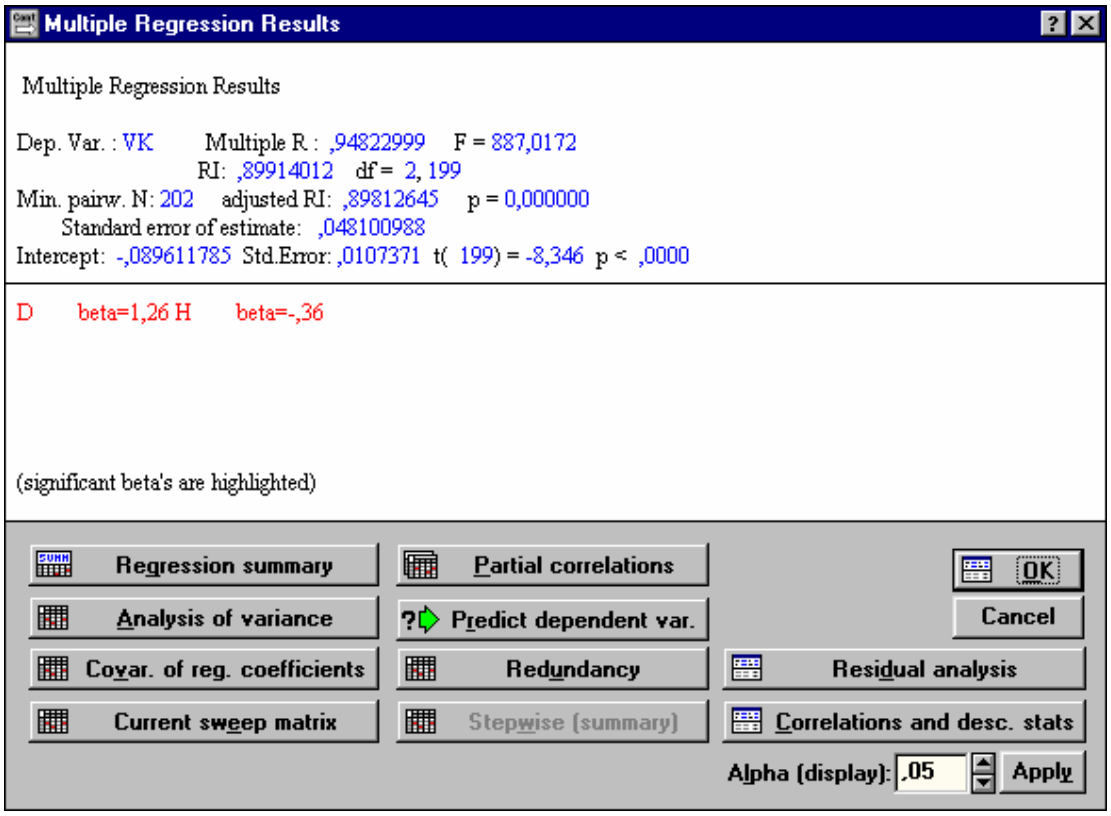


Рис. 33. Окно просмотра результатов регрессионного анализа

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

Multiple R - коэффициент множественной корреляции;

Характеризует тесноту линейной связи между зависимой и всеми независимыми переменными. Может принимать значения от 0 до 1.

R^2 или **RI** - коэффициент детерминации;

Численно выражает долю вариации зависимой переменной, объясненную с помощью регрессионного уравнения. Чем больше R^2, тем большую долю вариации объясняют переменные, включенные в модель.

adjusted R - скорректированный коэффициент множественной корреляции;

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

adjusted R^2 или **adjusted RI** - скорректированный коэффициент детерминации;

Скорректированный R^2 можно с большим успехом (по сравнению с R^2) применять для выбора наилучшего подмножества независимых переменных в регрессионном уравнении

F - F-критерий;

df - число степеней свободы для F-критерия;

p - вероятность нулевой гипотезы для F-критерия;

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

Intercept - свободный член уравнения;

Std.Error - стандартная ошибка свободного члена уравнения;

t - t-критерий для свободного члена уравнения;

p - вероятность нулевой гипотезы для свободного члена уравнения.

Beta - β-коэффициенты уравнения.

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

При помощи кнопок диалогового окна Multiple Regressions Results (рис. 33) результаты регрессионного анализа можно просмотреть более детально.

Кнопка **Regression summary** - позволяет просмотреть основные результаты регрессионного анализа (рис. 34): **BETA** - β-коэффициенты уравнения; **St. Err. of BETA** - стандартные ошибки β-коэффициентов; **B** - коэффициенты уравнения регрессии; **St. Err. of B** - стандартные ошибки коэффициентов уравнения регрессии; **t (95)** - t-критерии для коэффициентов уравнения регрессии; **p-level** - вероятность нулевой гипотезы для коэффициентов уравнения регрессии.

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

Continue...

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

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

Рис. 34. Основные результаты регрессионного анализа

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

Кнопка **Analysis of variance** - позволяет ознакомиться с результатами дисперсионного анализа уравнения регрессии (рис. 35). В строках таблицы дисперсионного анализа уравнения регрессии - источники вариации: Regress. - обусловленная регрессией, Residual- остаточная, Total - общая. В столбцах таблицы: Sums of Squares - сумма квадратов, df - число степеней свободы, Mean Squares - средний квадрат, F - значение F - критерия, p-level - вероятность нулевой гипотезы для F - критерия.

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

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

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

Рис.35 .Результаты дисперсионного анализа уравнения регрессии

Кнопка **Partial correlations** - позволяет просмотреть частные коэффициенты корреляции (Partial Cor.) между переменными (рис. 36). Частная корреляция - это корреляция между двумя переменными, когда одна или больше из оставшихся переменных удерживаются на постоянном уровне (т.е. имеют постоянное значение). Частные коэффициенты корреляции, как и парные, могут принимать значения от -1 до +1.

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

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

Рис. 36. Результаты расчета частных коэффициентов корреляции

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

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

Кнопка **Predict dependent var.** - позволяет рассчитать по полученному регрессионному уравнению значение зависимой переменной по значениям независимых переменных. На рис. 37 приводится пример расчета объема ствола дуба в коре при величине диаметра ствола - 14 см и высоты - 11 м. Предсказанный (Predictd) объем составил 0,1614 куб.м.

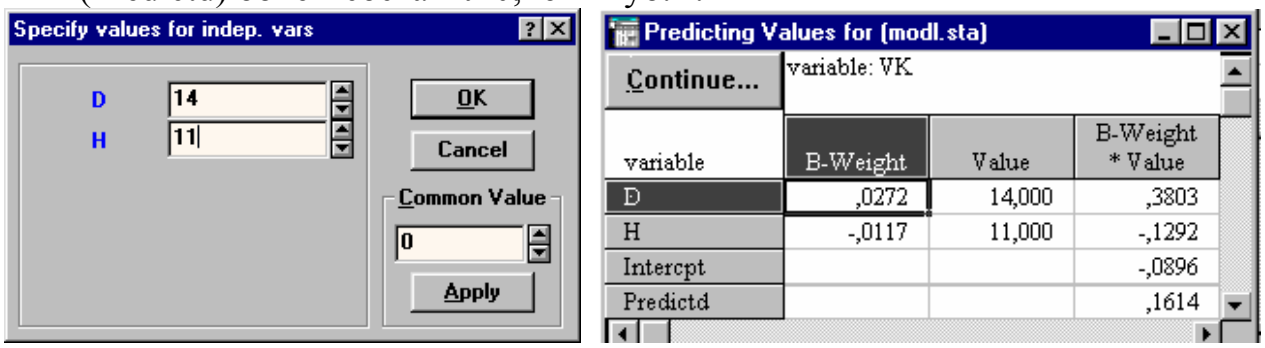


Рис. 37. Окно задания значений независимых переменных и результаты расчета по регрессионному уравнению зависимой переменной

Кнопка **Correlations and desc. stats** позволяет просмотреть описательные статистики и корреляционную матрицу с парными коэффициентами корреляции переменных, участвующих в регрессионной модели (рис. 38).

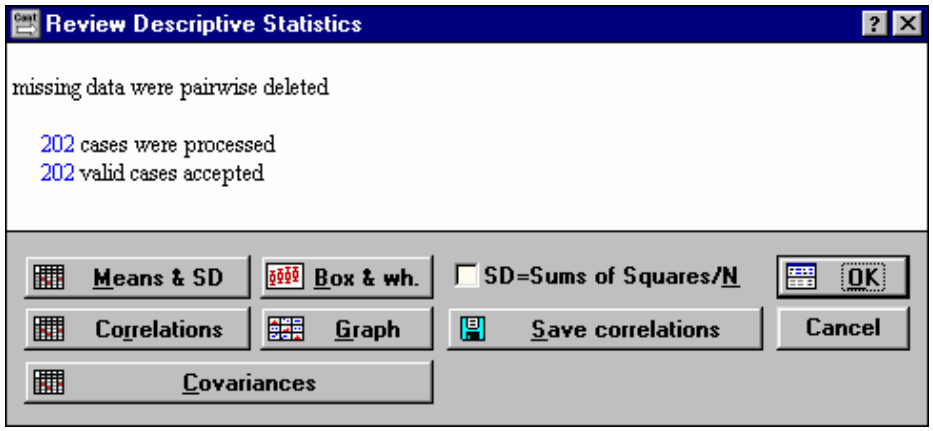


Рис. 38. Диалоговое окно Review Descriptive Statistics

Кнопка **Residual analysis** запускает процедуру всестороннего анализа остатков регрессионного уравнения (рис. 39). Остатки - это разности между опытными и предсказанными значениями зависимой переменной в построенной регрессионной модели.

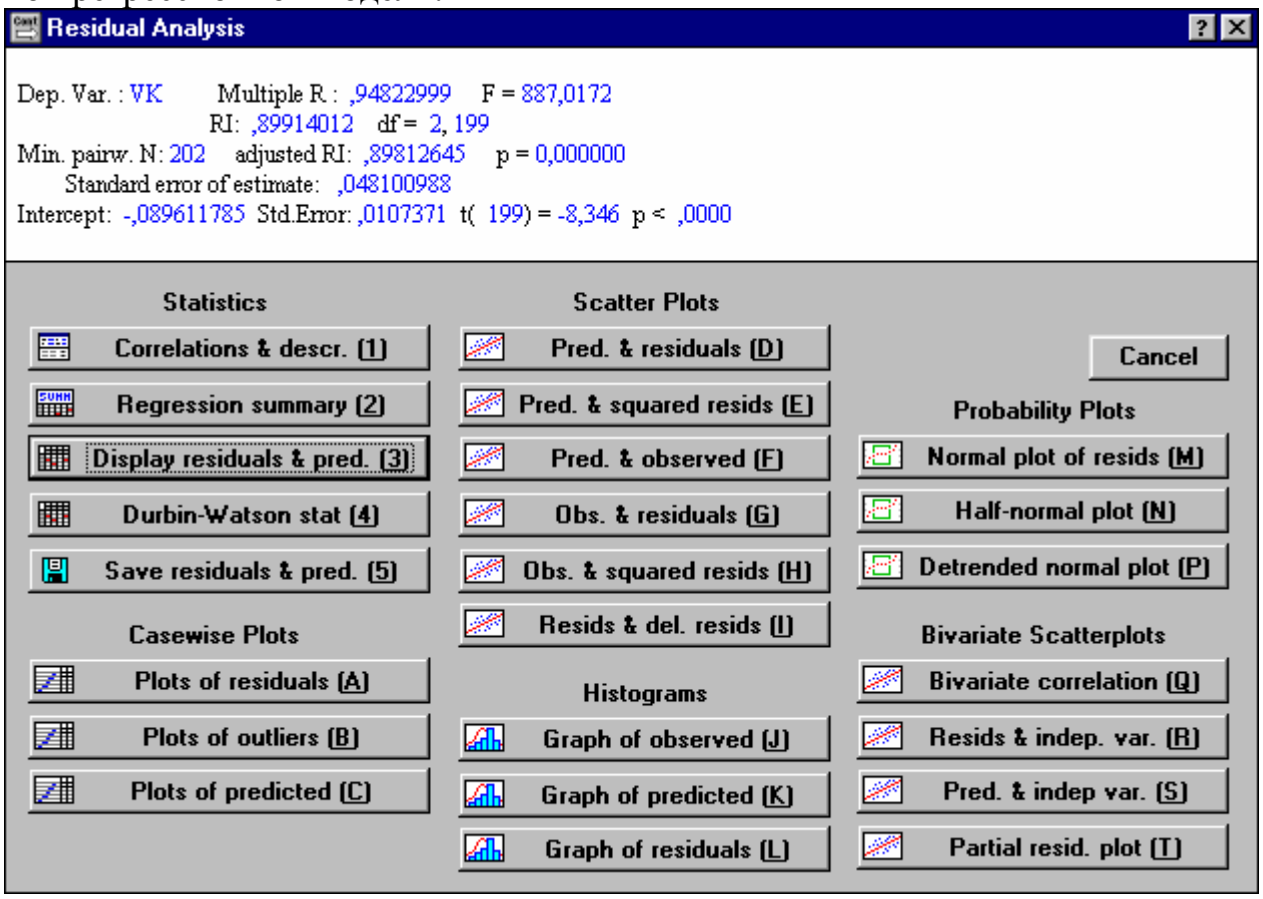


Рис.39 . Диалоговое окно Residual analysis (Анализ остатков)

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

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

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

Внимательный анализ остатков позволяет оценить адекватность модели. Остатки должны быть нормально распределены, со средним значением равным нулю и постоянной, независимо от величин зависимой и независимой перемен-

ных, дисперсией. Модель должна быть адекватна на всех отрезках интервала изменения зависимой переменной.

Просмотр величин остатков и специальных критериев, их оценивающих, осуществляется при помощи кнопки **Display residuals & pred.** окна Residual analysis. Для нашего примера фрагмент окна с этими данные представлен на рис. 40.

Predicted & Residual Values (modl.sta)

Continue... Dependent variable: VK

Case No.	Observed Value	Predictd Value	Residual	Standard Pred. v.	Standard Residual	Std.Err. Pred.Val	Mahalns. Distance	Deleted Residual	Cook's Distanc
202	,2148	,2241	-,0093	,348	-,193	,0050	1,147	-,0094	,000
203	,2375	,2517	-,0142	,541	-,294	,0043	,584	-,0143	,000
204	,3216	,3549	-,0333	1,263	-,692	,0057	1,850	-,0338	,001
Minimum	,0003	-,0753	-,1133	-1,747	-2,355	,0034	,003	-,1168	,000
Maximum	,6775	,5424	,2439	2,575	5,072	,0119	11,255	,2501	,22
Mean	,1744	,1744	-,0000	-,000	-,000	,0056	1,990	,0002	,00

Рис.40 . Окно со значениями остатков (Residuals), показателями Кука (Cook's Distance), расстояния Махаланобиса (Mahalns. Distance), опытными (Observed Value) и предстказанными по уравнению (Predictd Value) значениями зависимой переменной

Вполне достаточно бывает одного графического анализа остатков. О нормальности остатков можно судить по графику остатков на нормальной вероятностной бумаге. Чем ближе распределение к нормальному виду, тем лучше значения остатков ложатся на прямую линию. Он строится при помощи кнопки **Normal plot of resids.** окна Residual analysis (рис. 41).

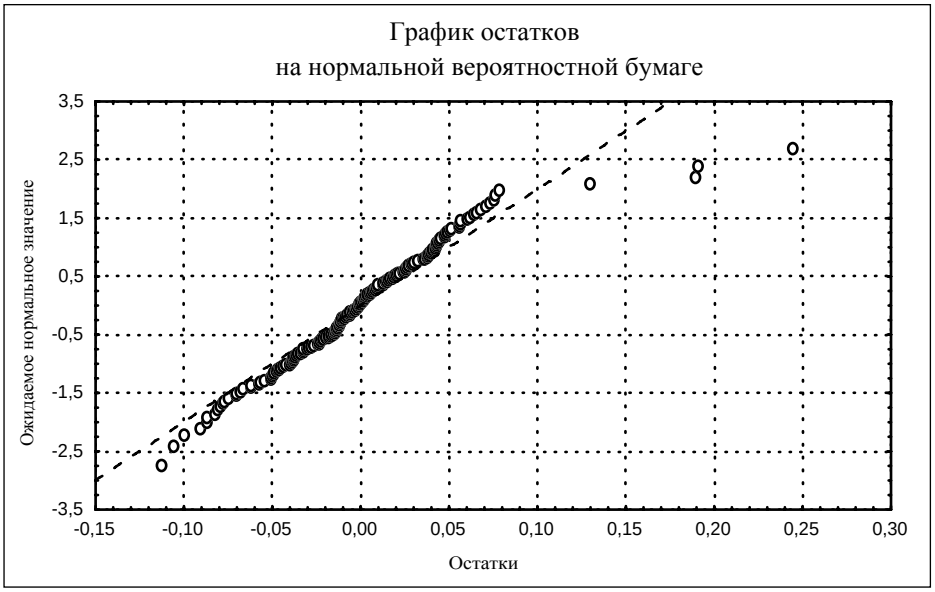


Рис..41. График остатков на нормальной вероятностной бумаге

Важно просмотреть графики зависимости остаток от каждой из независимых переменных. Их легко просмотреть при помощи кнопки **Resids & indep.**

var. окна Residual analysis. Остатки должны быть нормально распределены, т.е. на графике они должны представлять приблизительно горизонтальную полосу одинаковой ширины на всем ее протяжении. Коэффициент корреляции (r) между регрессионными остатками и переменными должен равняться нулю.

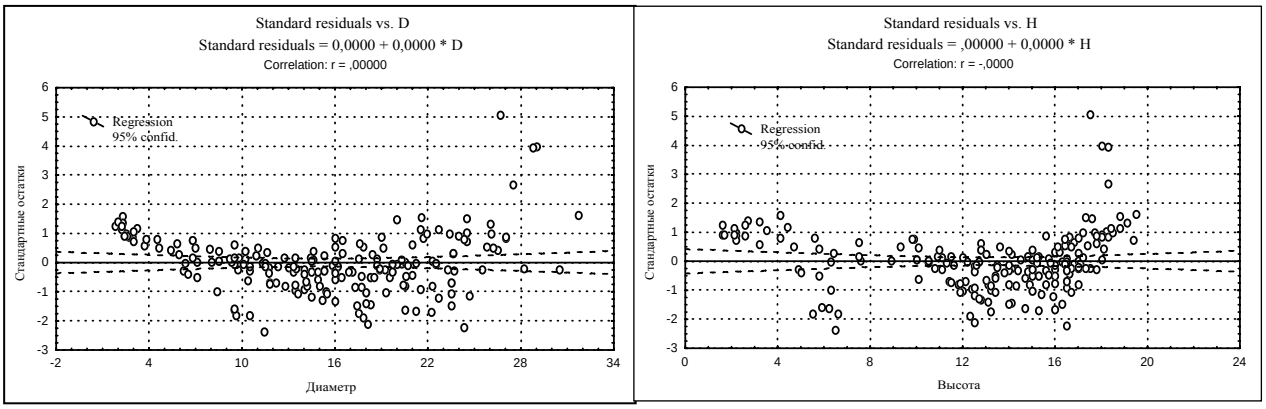


Рис. 42. Зависимость остатков от независимых переменных: диаметра и высоты

В нашем случае на графиках остатков (рис. 42) хорошо просматривается нелинейный тренд, что вызывает сомнение в адекватности модели. Присутствие нелинейного тренда в регрессионных остатках говорит о необходимости пересмотра модели (преобразования или ввода новых переменных, перехода от линейной модели к нелинейной).

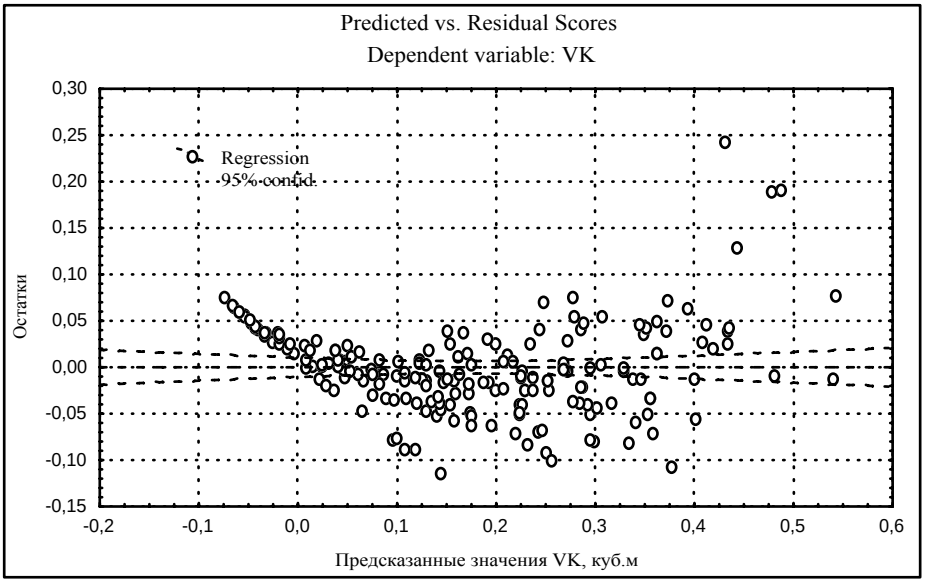


Рис. 43. Зависимость регрессионных остатков от предсказанных значений зависимой переменной

Для выявления нестабильности дисперсии ошибки уравнения при помощи кнопки **Pred. & residuals** окна Residual analysis можно создать график зависимости регрессионных остатков от предсказанного значения зависимой

переменной. Рис. 43. позволяет заключить о непостоянстве дисперсии ошибки уравнения (с увеличением значений зависимой переменной дисперсия увеличивается). Это еще одной подтверждение неадекватности анализируемой модели.

Очень удобным визуальным способом оценки адекватности регрессионной модели является анализ графического изображения опытных и полученных по регрессионному уравнению значений зависимой переменной. Оно строится при помощи кнопки **Pred. & observed** окна Residual analysis.

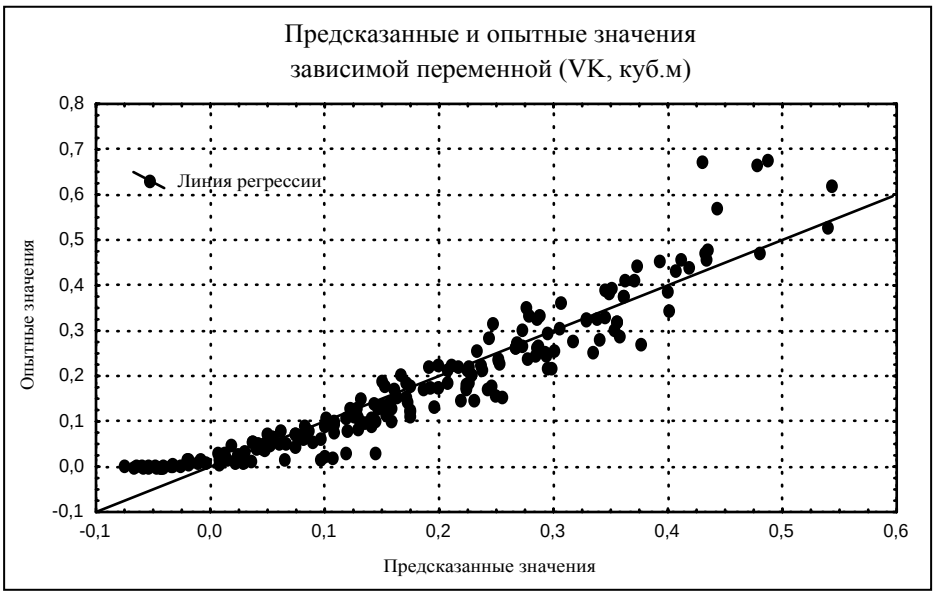


Рис.44 . Линия регрессии, опытные и полученные по регрессионному уравнению значений зависимой переменной

Из рис. 44 хорошо видно, что линейный вид нашей модели плохо описывает взаимосвязь объема ствола дуба в коре от его диаметра и высоты (модель при малых и больших значениях отклика занижает величину зависимой переменной). Эта связь носит нелинейный характер.

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

Опцию **Mode** стартового окна регрессионного анализа (рис. 27) выставим в положение **Fixed non linear**.

Если выбран фиксированный нелинейный тип регрессионной модели, то после нажатия на кнопку ОК в диалоговом окне Multiple Regressions (рис. 45), появляется окно Non-linear Components Regression (рис. .), в котором можно выбрать следующие типы преобразования переменных: X^2, X^3, X^4, X^5, $\sqrt{X}$ ($X \geq 0$), $\ln X$ ($X > 0$), $\lg_{10} X$ ($X > 0$), e^X ($40<X<-40$), 10^X (-18 to +18), $1/X$ ($X \neq 0$). Если потребуются какие либо иные преобразования переменных, то тогда в файле данных следует

создать мнимые вичисляемые переменные и включить их в качестве зависимых переменных в регрессионную модель.

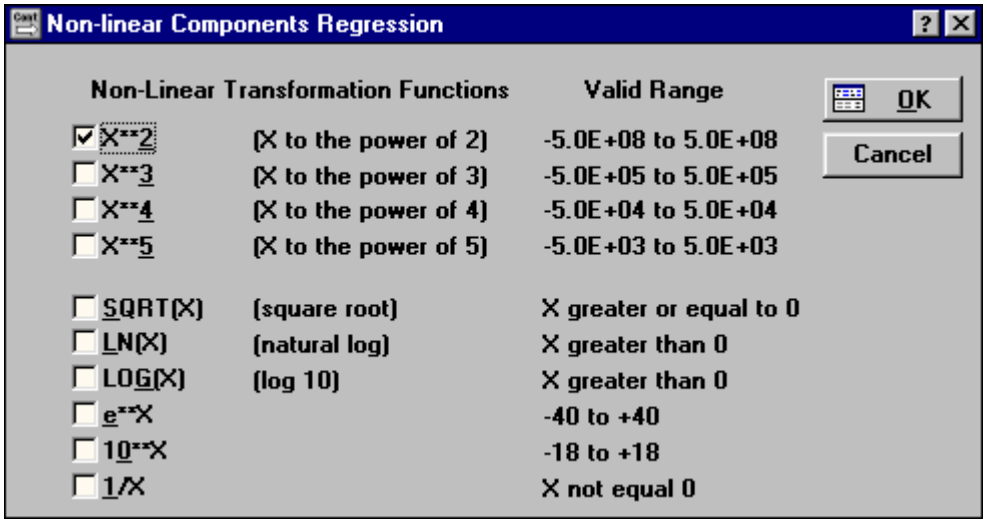


Рис. 45.Окно выбора типов преобразования переменных

После того, как тип преобразования переменных определен (в нашем примере это возведение в квадрат), необходимо уточнение зависимой и независимых переменных фиксированной нелинейной регрессионной модели. Оно производится на следующем шаге при помощи кнопки Variables диалогового окна Model Definition (Уточнение модели) (рис. 46).

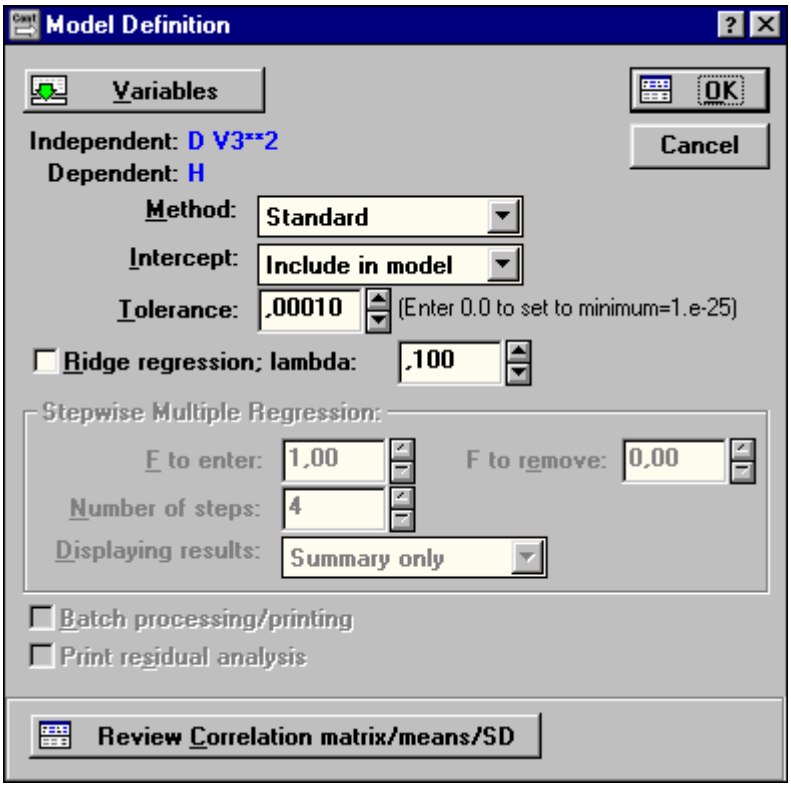


Рис.46. Диалоговое окно Model Definition (Уточнение модели)

Зависимой (dependent) переменной в нашем случае будет - VK; независимыми (independent) - D и D^2 (рис. 47). Переменная D^2 значится в списке переменных как V3**2, так как переменная D является третьей в списке переменных.

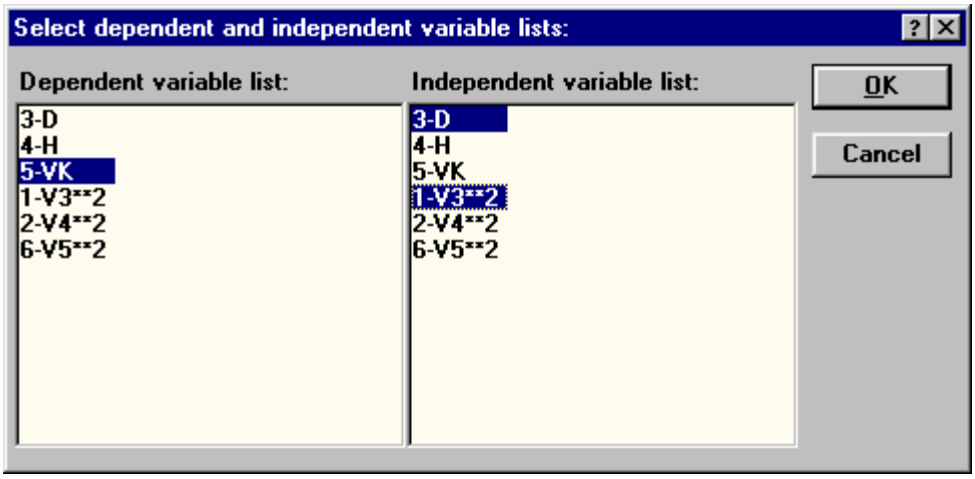


Рис. 47. Выбор переменных для расчета уравнения $VK = a_1 + a_2D + a_3D^2$

Уравнение взаимосвязи между объемом ствола дуба в коре (VK) от его диаметром (D) оказалось сследующее: VK = 0,00023 - 0,0034D + 0,0008D^2. Все коэффициенты уравнения (за исключением свободного члена) значимы на 5% уровне (p-level < 0,05). Это уравнение объясняет 95,8% (R^2 = 0,958) вариации зависимой переменной (рис. 48).

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

Continue...

R= ,97900982 RI= ,95846023 Adjusted RI= ,95804274
F(2,199)=2295,8 p<0,0000 Std.Error of estimate: ,03087

N=202	BETA	St. Err. of BETA	B	St. Err. of B	t(199)	p-level
Intercpt			,0023	,0087	,268	,7891
D	-,1561	,0568	-,0034	,0012	-2,750	,0065
V3**2	1,1292	,0568	,0008	,0000	19,888	0,0000

Рис. 48. Результаты регрессионного анализа модели $VK = a_1 + a_2D + a_3D^2$

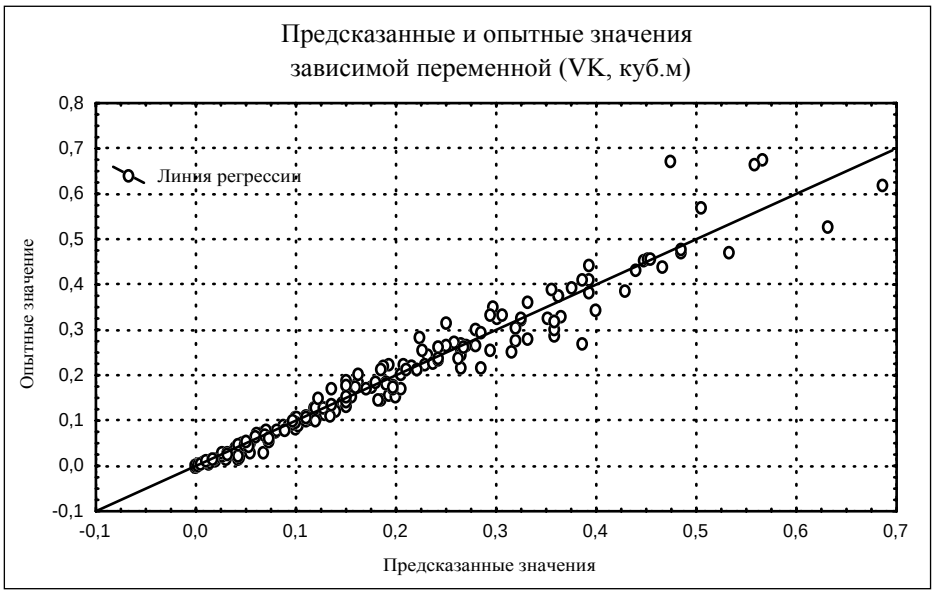


Рис.49. Линия регрессии, опытные и полученные по регрессионному уравнению значений зависимой переменной

По всем стандартным параметрам второе уравнение регрессии значительно лучше первого. Это наглядно подтверждает и график на рис. 49.

Найдем параметры еще одного регрессионного уравнения. Вид уравнения: $VK = a_1 D^{a2} H^{a3}$. Это степенное уравнение может быть приведено к линейному виду через логарифмирование: $\ln VK = \ln a_1 + a_2 \ln D + a_3 \ln H$.

При помощи кнопки Variables укажем зависимую - VK и независимые переменные - D,H. Опцию Mode стартового окна регрессионного анализа (рис. 27) выставим в положение Fixed non linear. В качестве типа преобразования переменных выберем натуральный логарифм (ln (X)). В диалоговом окна Model Definition при помощи кнопки Variables уточним модель, переопределив зависимую и независимые переменные так, как это показано на рис. 50.

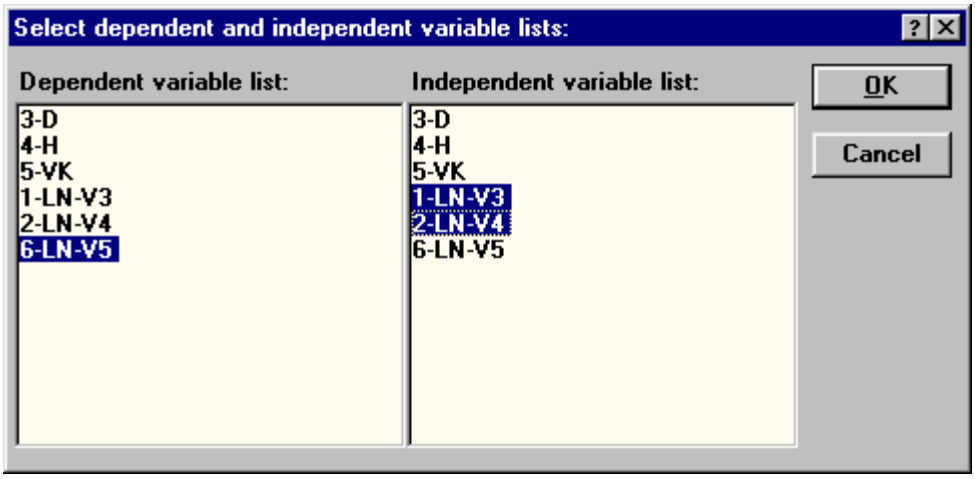
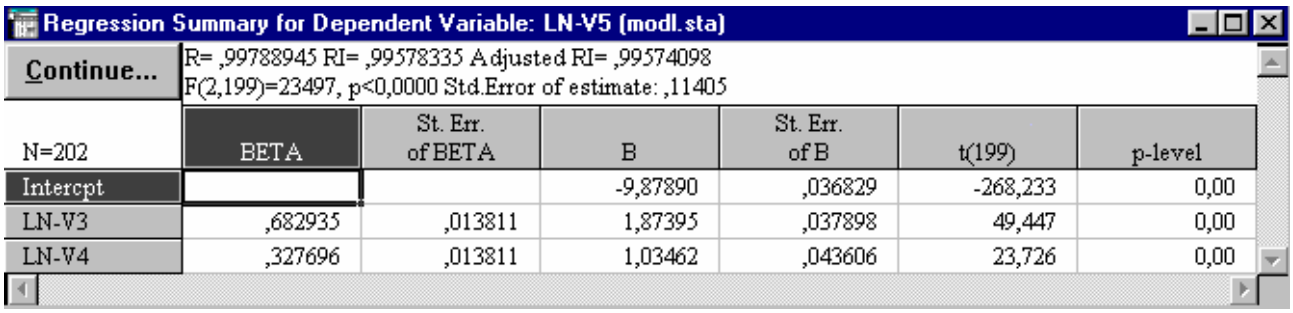


Рис. 50. Выбор переменных для расчета уравнения $lnVK = lna_1 + a_2 lnD + a_3 lnH$

Основные результаты регрессионного анализа представлены на рис. 51.



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

Рис. 51. Результаты регрессионного анализа модели $lnVK = lna_1 + a_2 lnD + a_3 lnH$

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

Проверим адекватность полученной модели через анализ остатков. В целом он даст положительное заключение. В качестве иллюстрации приведем лишь несколько графиков (рис. 52, 53), подтверждающих такой вывод.

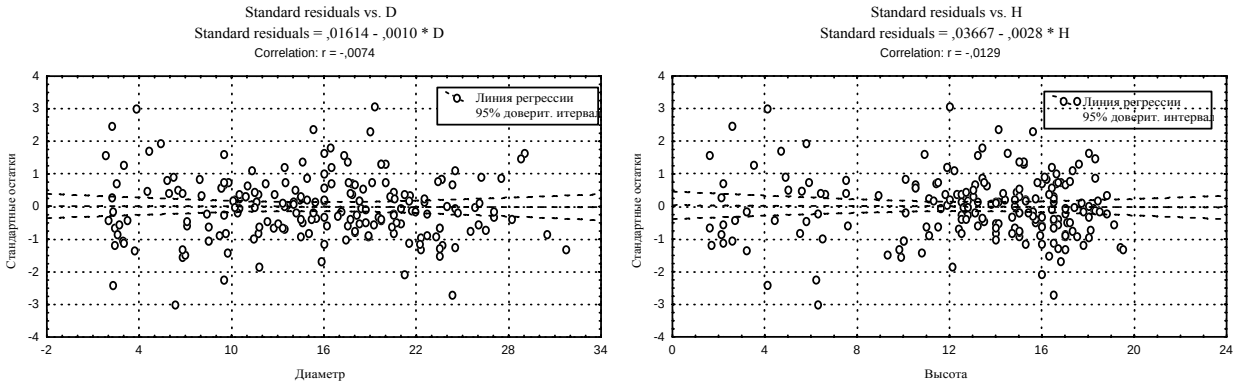


Рис. 52. Зависимость остатков степенного уравнения от независимых переменных: диаметра и высоты

Рис.53. Линия регрессии, опытные и полученные по степенному регрессионному уравнению значений зависимой переменной

Поиск наилучшей регрессионной модели представляет собой довольно громоздкий процесс. При помощи опции **Method** (рис. 27) пользователь может отказаться от стандартного проведения регрессионного анализа (Standard) и воспользоваться методами пошагового включения переменных в регрессионную модель (**Forward stepwise**) или пошагового исключения переменных (**Backward stepwise**) из регрессионной модели. Опция **Displaying results** позволяет просматривать или же только итоговые результаты регрессионного анализа (Summary only) или после каждого шага включения или исключения переменных (At each step). Если необходимо получить регрессионную модель без свободного члена уравнения, тогда в списке поля **Intercept** нужно выбрать - Set to zero.

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

В начале создадим новую переменную - DH. В файле данных она будет одиннадцатой по счету. Для расчета значений этой переменной вызовем окно с экспликацией этой переменной (рис. 54) и в поле **Long name** введем формулу, в соответствии с которой значения переменной должны быть рассчитаны, т.е "=V3*V4".

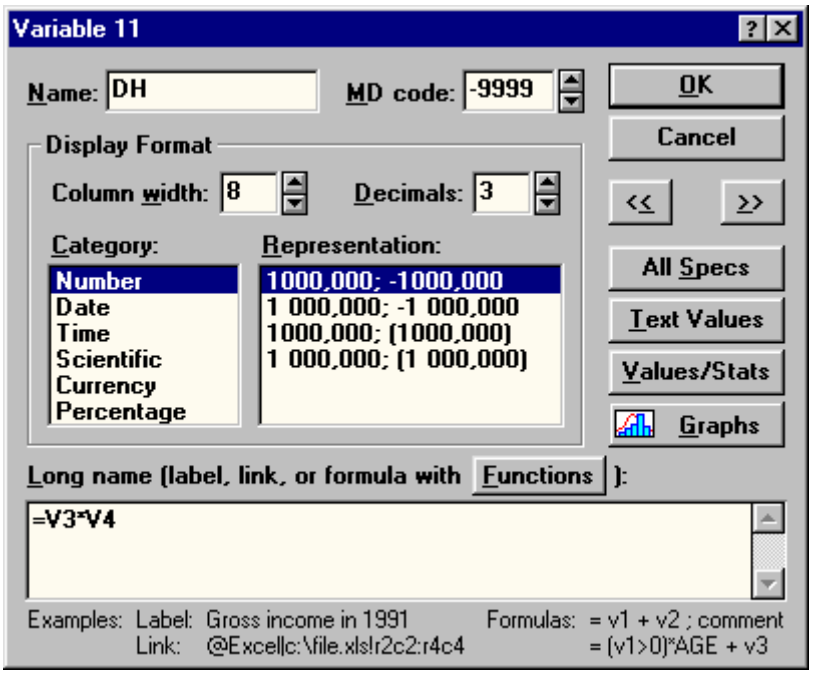


Рис.54. Окно экспликации 11-ой переменной

Опцию **Mode** стартового окна регрессионного анализа (рис.27) выставим в положение **Fixed non linear**.

Определим тип преобразования переменных - возведение в квадрат (рис. 45) и уточним зависимую и независимые переменные модели (рис. 55).

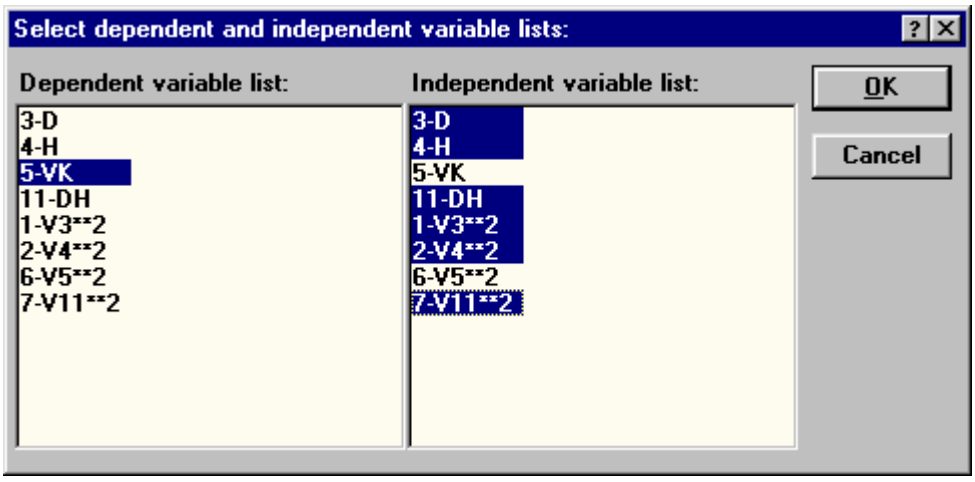


Рис.55. Уточнение зависимой и независимых переменных регрессионного анализа

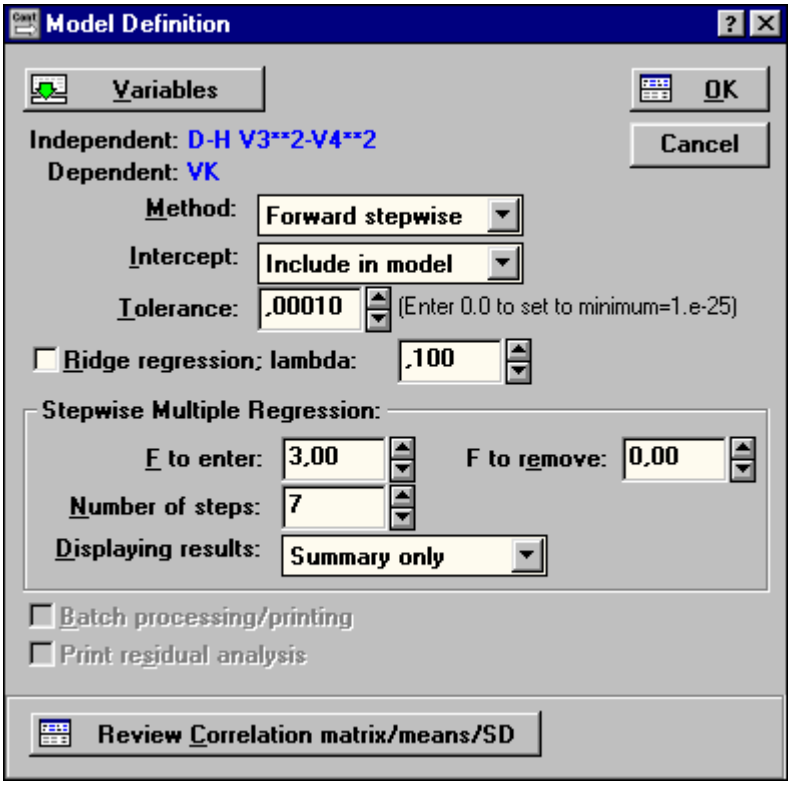


Рис.56. Диалоговое окно Model Definition при использовании метода пошагового включения переменных в модель

Для пошаговых методов регрессионного анализа важно установить величину **Tolerance** (толерантность) и величины частного F- критерия для включения в модель (**F to enter**) и исключения из нее (**F to remove**). Установив величину толерантности мы создаем барьер для включения в модель переменных, толерантность которых меньше установленной. Если величина толерантности переменной мала, то переменная несет малую дополнительную информацию и включение ее в модель не целесообразно. Какая либо новая независимая переменная, включаемая в модель, может сильно влиять на зависимую переменную, но если она включается в модель после других переменных, она может уже мало влиять на переменную отклика (например, из-за сильной коррелированности с переменными, уже включенными в модель). По умолчанию в пакете Statistica переменная включается в модель, если частный F- критерий больше или равен 1. Численное значение F- критерия для включения никогда не выбирается меньшим, чем численное значение F- критерия для исключения.

Выставим опции окна Model Definition так, как показано на рис. 56. В результате процедуры пошагового включения переменных в регрессионную модель получено следующее уравнение (рис.): $VK = 0{,}0214 + 0{,}0009D^2 - 0{,}0104D + 0{,}0003(DH)^2$. Все коэффициенты уравнения значимы на 5% уровне (p-level < 0,05). Это уравнение объясняет 96,4% ($R^2 = 0{,}964$) вариации зависимой переменной (рис. 57). Средняя ошибка уравнения составляет 0,02862 $м^3$.

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

Continue...

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

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

Рис.57. Характеристика уравнения, полученного методом Forward stepwise

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

1. Регрессионная модель должна объяснять не менее 80% вариации зависимой переменной, т.е. $R^2 \geq 0.8$.
2. Стандартная ошибка оценки зависимой переменной по уравнению должна составлять не более 5% среднего значения зависимой переменной;
3. Коэффициенты уравнения регрессии и его свободный член должны быть значимы на 5%-ом уровне.
4. Остатки от регрессии должны быть без заметной автокорреляции ($r<0,30$), нормально распределены и без систематической составляющей.

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

Одним из недостатков классического регрессионного анализа, в основе которого лежит метода наименьших квадратов, является недостаточная устойчивость к изменениям входной информации. Сейчас довольно широко стали применяться альтернативные регрессионные модели, одной из которых является **гребневая регрессия,** которая отличается устойчивостью для случаев сильной коррелированности зависимых переменных друг с другом. В отличии от метода наименьших квадратов, дающего несмещенные оценки коэффициентов уравнения, в методе гребневой регрессии оценки смещенные, но при этом они имеют меньшую дисперсию. Поэтому такие оценки могут давать более точные и приемлемые для практического использования модели (Забелин, 1983).

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

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

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

СПИСОК ЛИТЕРАТУРЫ

1. Боровиков В.П. Популярное введение в программу Statistica. М.: КомпьютерПресс, 1998.- 267с.
2. Боровиков В.П., Боровиков И.П. Statistica. Статистический анализ и обработка данных в среде Windows. -М.: Информационно-издательский дом "Филинъ", 1997.- 608с.
3. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. Кн. 1, 2. М.: Мир, 1981.- 252с.
4. Лакин Г.Ф. Биометрия.- М.: Высшая школа, 1990.- 352 с.
5. Литтл Т., Хиллз Ф. Сельскохозяйственное опытное дело. Планирование и анализ. - М.: Колос, 1981.- 320с.
6. Никитин К.Е., Швиденко А.З. Методы и техника обработки лесоводственно-таксационной информации..- М.:Лесная промышленность, 1978.- 272с.
7. Тюрин Ю.П., Макаров А.А. Анализ данных на компьютере.- М.: ИНФРА-М, Финансы и статистика, 1995.- 384с.
8. Уланова Е.С., Забелин В.Н. Методы корреляционного и регрессионного анализа в агрометеорологии.- Л.: Гидрометеоиздат, 1990.- 207с.