ФГБОУ ВО «Воронежский государственный технический

университет»

Кафедра системного анализа и управления
в медицинских системах

МЕТОДИЧЕСКИЕ УКАЗАНИЯ

к выполнению лабораторных работ по дисциплине
"Автоматизация обработки биомедицинской информации"
для студентов направления 12.03.04
"Биотехнические системы и технологии"
(профили «Биотехнические и медицинские аппараты и системы»,
«Менеджмент и управление качеством в здравоохранении»)
очной, заочной форм обучения

Воронеж 2016

Составитель: д-р техн. наук Е.Н. Коровин

УДК 681.327.8

Методические указания к выполнению лабораторных работ по дисциплине "Автоматизация обработки биомедицинской информации" для студентов направления 12.03.04 «Биотехнические системы и технологии» (профили «Биотехнические и медицинские аппараты и системы», «Менеджмент и управление качеством в здравоохранении») очной и заочной форм обучения / ФГБОУ ВО «Воронежский государственный технический университет»; сост. Е.Н. Коровин. Воронеж, 2018. 36 с.

Данные методические указания предназначены для выполнения лабораторных работ по дисциплине «Автоматизация обработки биомедицинской информации».

Предназначены для студентов 3 курса.

Рецензент д-р техн. наук, проф. И.Я. Львович

Ответственный за выпускзав. кафедрой
д-р техн. наук, проф. О.В. Родионов

Печатается по решению редакционно-издательского совета Воронежского государственного технического университета

ЛАБОРАТОРНАЯ РАБОТА № 1

Корреляционный анализ

Цель работы: Выявить корреляционные связи между переменными

1. Общие сведения

Корреляция в математической статистике - это вероятностная или статистическая зависимость, не имеющая строго функционального характера. В отличие от функциональной корреляционная зависимость возникает, когда один из показателей (признаков) зависит не только от второго показателя, но и от ряда случайных факторов.

Корреляционный анализ представляет собой определение зависимости, тесноты связи между двумя случайными величинами (признаками или факторами) и включает в себя построение корреляционного поля и составление корреляционной таблицы, вычисление выборочных коэффициентов корреляции.

Для того, чтобы знать в каком направлении влияет одна случайная переменная *x* на другую *y*, т.е. оценить, как сильно они связаны, используется коэффициент парной корреляции R_{xy}, асимптотически несмещенная оценка которого

$$R_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}, \qquad (1)$$

где x_i - *i*-е значение переменной x; y_i - *i*-е значение переменной y; N - общее число наблюдений; $\bar{x}$ - среднее значение переменной x из выборочной совокупности; $\bar{y}$ - среднее значение переменной y из

выборочной совокупности; s_x- выборочная оценка среднеквадратического отклонения переменной x; s_y- выборочная оценка среднеквадратического отклонения переменной y

$$s_x = \sqrt{s_x^2}\ ;\ s_y = \sqrt{s_y^2}\ , \qquad (2)$$

где s_x^2 и s_y^2 - соответственно выборочные дисперсии переменных x и y.

Коэффициент парной корреляции является показателем тесноты и направления корреляционной связи двух случайных переменных, и его значение находится в пределах $-1 \le R_{xy} \le +1$.

При отсутствии корреляционной связи между двумя случайными переменными коэффициент парной корреляции $R_{xy} = 0$, в этом случае корреляционная связь между переменными x и y отсутствует. Если связь между двумя переменными линейная и функциональная, тогда $R_{xy} = +1$ или $R_{xy} = -1$.

Геометрической интерпретацией коэффициента парной корреляции является поле корреляции, представляющее собой двумерное поле рассеивания (рис.1).

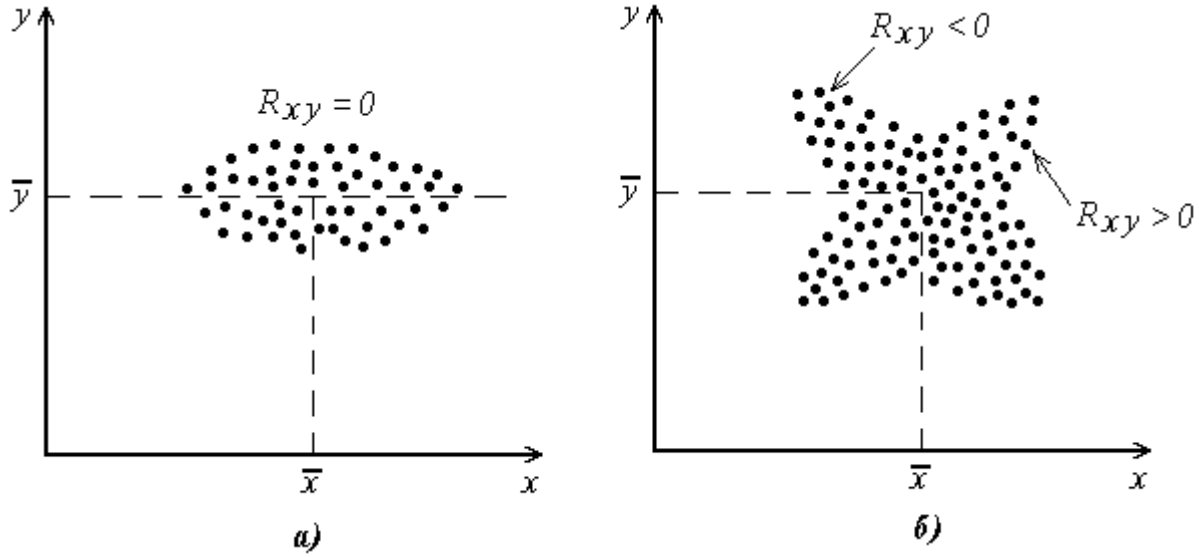


Рис.1. Геометрическая интерпретация коэффициента парной корреляции: а) R_{xy}=0; б) $R_{xy} \neq 0$

Если расчетное значение коэффициента парной корреляции $R_{xy_{\text{р}}}$ окажется больше критического значения $R_{xy_{\text{кр}}}$, найденного по статистической таблице (приложение), то гипотеза о статистической значимости тесноты корреляционной связи при числе степеней свободы f=N-2, где N - объем выборки, при заданном уровне значимости q % принимается.

Пример 1. Требуется определить коэффициент парной корреляции между случайными величинами x и y по выборке объемом N=25, полученной из независимых опытов (табл.1).

Таблица 1

Переменные	Номер опыта												
	1	2	3	4	5	6	7	8	9	10	11	12	13
x	15	22	18	15	24	22	26	23	25	21	20	19	27
$\tilde{y}$	13.5	16.5	14	13.5	19.5	18.5	20.5	19	20	17	17	16	25

Продолжение табл. 1

Переменные	14	15	16	17	18	19	20	21	22	23	24	25
x	26	25	23	18	13	19	13	25	24	28	25	25
$\tilde{y}$	24	23	14.2	11.8	15.2	12.5	23.6	19.6	25.3	20.9	21	24

Находим оценки математического ожидания:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{541}{25} = 21.64;$$

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{465.1}{25} = 18.6.$$

Рассчитываем вспомогательные значения $(x_i - \bar{x})$, $(y_i - \bar{y})$, $(x_i - \bar{x})^2$, $(y_i - \bar{y})^2$, $(x_i - \bar{x})(y_i - \bar{x})$, $(i = \overline{1, N})$ и записываем их в табл. 2.

Таблица 2

Номер опыта	x	y	Вспомогательные величины				
			$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})$ x x $(y_i - \bar{y})$
1	15	13,5	-6,64	-5,1	44,09	26,01	33,86
2	22	16,5	0,36	-2,1	0,13	4,41	-0,76
3	18	14	-3,64	-4,6	13,25	21,16	16,74
4	15	13,5	-6,64	-5,1	44,09	26,01	33,86
5	24	19,5	2,36	0,9	5,57	0,81	2,12
6	22	18,5	0,36	-0,1	0,13	0,01	-0,036
7	26	20,5	4,36	1,9	19,01	3,61	8,24
8	23	19	1,36	0,4	1,85	0,16	0,54
9	25	20	3,36	1,4	11,29	1,96	4,70
10	21	17	-0,64	-1,6	0,41	2,56	1,02
11	20	17	-1,64	-1,6	2,69	2,56	2,62
12	19	16	-2,64	-2,6	6,97	6,76	6,86
13	27	25	5,36	6,4	28,76	40,96	34,30
14	26	24	4,36	5,4	19,01	29,16	23,54
15	25	23	3,36	4,4	11,29	19,36	14,78
16	23	14,2	1,36	-4,4	1,85	19,36	-5,98
17	18	11,8	-3,64	-6,8	13,25	46,24	24,75
18	13	15,2	-8,64	-3,4	74,65	11,56	29,38
19	19	12,5	-2,64	-6,1	6,97	37,21	16,10
20	13	23,6	-8,64	5,0	74,65	25	-43,20
21	25	19,6	3,36	1,0	11,29	1,0	3,36
22	24	25,3	2,36	6,7	5,57	44,89	15,81
23	28	20,9	6,36	2,3	40,45	5,29	14,63
24	25	21	3,36	2,4	11,29	5,76	8,06
25	25	24	3,36	5,4	11,29	29,16	18,14
∑	541	465,1	-	-	459,89	410,97	313,41
-	$\bar{x}$	$\bar{y}$	s_x^2	s_y^2	s_x	s_y	R_{xy}
-	21,64	18,6	19,16	17,12	4,377	4,138	0,721

Определяем оценки дисперсий переменных x и y и их

среднеквадратические отклонения:

$$s_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{459.89}{25-1} = \frac{459.89}{24} = 19.16;$$

$$s_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2 = \frac{410.97}{24} = 17.12;$$

$$s_x = \sqrt{s_x^2} = \sqrt{19.16} = 4.377;$$

$$s_y = \sqrt{s_y^2} = \sqrt{17.12} = 4.138.$$

Коэффициент парной корреляции между переменными x и y оцениваем по (1)

$$R_{xy_{\text{р}}} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y} = \frac{313{,}41}{24 \cdot 4{,}377 \cdot 4{,}138} = 0{,}721.$$

Так как $R_{xy_{\text{р}}}$=0,721 оказалось больше критического значения $R_{xy_{\text{кр}}}$=0,423 при f=N-2=23 и q=5 %, то гипотеза о наличии корреляционной связи между случайными величинами x и y принимается, а полученная из наблюдений величина R_{xy} значимо отличается от нуля и наличие линейной корреляции не вызывает сомнения.

Коэффициент корреляции (1) характеризует степень тесноты линейной зависимости между случайными величинами.

В случае изучения зависимости случайной величины x от случайных величин y и z используется коэффициент множественной корреляции:

$$R_{x(yz)} = \sqrt{\frac{r_{xy}^2 + r_{xz}^2 - 2r_{xy}r_{xz}r_{yz}}{1 - r_{yz}^2}}, \qquad (3)$$

где r_{xy}, r_{xz}, r_{yz} - парные коэффициенты корреляции, а $0 \le R_{x(yz)} \le 1$.

Использование коэффициентов парной и множественной корреляции неявно предполагает нормальное распределение

генеральных совокупностей, из которых производится выборка.

Как отмечалось выше, коэффициент корреляции характеризует степень линейной зависимости между переменными, при этом нелинейная связь не обнаруживается. В этом случае для оценки используется корреляционное отношение. Корреляционное отношение y и x называется отношение межгруппового среднего квадратичного отклонения δ_y переменной y к ее общему среднему квадратичному отклонению σ_y

$$\eta_{yx} = \frac{\delta_y}{\sigma_y}, \tag{4}$$

где межгрупповая дисперсия определяется по формуле

$$\delta_y = \frac{\sum_{i=1}^{n} (\bar{y}_i - \bar{y})^2 \cdot n_i}{n}. \tag{5}$$

Аналогично определяется корреляционное отношение x и y η_{xy}. Основные свойства корреляционных отношений:

1) $0 \le \eta_{yx} \le 1$, $0 \le \eta_{xy} \le 1$;

2) если $\eta = 0$, то корреляционная связь отсутствует;

3) если $\eta = 1$, то переменные связаны функционально;

4) для линейной зависимости между переменными x и y необходимо и достаточно, чтобы выполнялось равенство $|r| = \eta_{yx}$;

5) $\eta_{yx} \neq \eta_{xy}$;

6) $0 \le |r| \le \eta \le 1$.

В других случаях (когда вид распределения неизвестен) используются меры связи, не регламентирующие нормальность выборок (методы непараметрической статистики), например, коэффициент ранговой корреляции Спирмена - r_s:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \qquad (6)$$

где d_i^2 - квадраты разности между рангами сопряженных признаков, n – число наблюдений (число пар рангов).

При полной связи ранги признаков совпадают (d=0) и следовательно r_s=1.

Статистическая значимость коэффициента ранговой корреляции оценивается с помощью следующей статистики:

$$tr_s = \frac{z_q}{\sqrt{n-1}}\left(1 - \frac{m}{n-1}\right), \qquad (7)$$

где z_α и m связаны соотношениями с уровнем значимости: для q = 5 % z=1,96 и m=0,16; для q = 1 % z=2,58 и m=0,69. Нулевая гипотеза отвергается, если полученное значение r_s превзойдет или окажется равным рассчитанному критическому значению tr_s.

Пример 2. Требуется оценить зависимость между содержанием вещества B в ткани C (X_1) и приростом концентрации вещества D в крови (X_2) у 10 пациентов, получавших препарат A (табл. 3) на основе коэффициента ранговой корреляции Спирмена.

Таблица 3

Пациент	Параметр X_1	Параметр X_2	Ранг R_{X1}	Ранг R_{X2}	d_i	d_i^2
1	8	4	4,5	5	-0,5	0,25
2	8	5	4,5	8,5	-4,0	16,0
3	9	4	7	5	2,0	4,0
4	10	3,5	9	2,5	6,5	42,25
5	7	5	2,5	8,5	-6,0	36,0
6	7	5	2,5	8,5	-6,0	36,0
7	9	3,5	7	2,5	4,5	20,25
8	9	4	7	5	2,0	4,0
9	11	2	10	1	9,0	81,0
10	6	5	1	8,5	-7,5	56,25
						$\sum_{i=1}^{n} d_i^2$ =296

Если бы отдельные варианты ряда не повторялись, их рангами были бы натуральные числа от 1 в порядке возрастания. Но одинаковым значениям вариант присваиваются ранги, равные средним арифметическим их рангов. Величина d_i представляет собой попарные разности рангов изучаемых выборок. В качестве правила для проверки правильности ранжирования используют равенство 0 суммы d_i .

По формуле (6) для n=10 получаем ранговый коэффициент корреляции r_s= -0,79. Критическое значение tr_s рассчитанное по формуле (1.7) для уровня значимости 5 % (z=1,96; m=0,16) равно 0,64. Так как значение рангового коэффициента корреляции по модулю превосходит соответствующее критическое значение, с вероятностью 95 % можно утверждать, что между сравниваемыми параметрами существует значимая отрицательная корреляционная связь.

Пример 3. Рассмотрим построение корреляционной матрицы на примере анализа показателей деятельности лечебно-профилактического учреждения. Мы так же проведем графический анализ полученной матрицы. В качестве исходных данных будем использовать данные из файла *DATA.STA.*

Имеется система переменных Y_1...Y_3, X_4…X_{17}. Рассматриваются следующие показатели:

Y 1 – производительность труда врачей;

Y 2 – индекс снижения себестоимости оказания медицинских услуг;

Y 3 – рентабельность;

X 4 – трудоемкость оказания единицы медицинской услуги;

X 5 – удельный вес младшего медицинского персонала в составе медицинского персонала;

X 6 – удельный вес медицинских услуг;

X 7 – коэффициент сменности диагностического оборудования;

X 8 – премии и вознаграждения на одного работника ЛПУ;

X 9 – удельный вес потерь от некачественной медицинской услуги;

X 10 – фондоотдача ЛПУ;

X 11 – среднегодовая численность медицинского персонала;

X 12 – среднегодовая стоимость ОПФ ЛПУ;

X 13 – среднегодовой фонд заработной платы персонала ЛПУ;

X 14 – фондовооруженность труда ЛПУ;

X 15 – оборачиваемость нормированных оборотных средств ЛПУ;

X 16 – оборачиваемость ненормированных оборотных средств ЛПУ;

X 17 – непроизводственные расходы ЛПУ.

В модуле **Основные статистики (Basic Statistics/ Tables)** легко можно вычислить и проанализировать корреляционную матрицу выбранных вами переменных. Для начала проведем корреляционный анализ переменных *Y*1, *X*4, *X*5 и *X*6.

• Запустите программу **STATISTICA**. Переключитесь в модуль **Основные статистики (Basic Statistics/ Tables)**. Нажмите кнопку **Open data (Открыть файл данных)** и откройте файл *data.sta.*

•Принимаем априори, что исходные выборки нормально распределены, поэтому для установления тесноты статистической связи можем воспользоваться параметрическим коэффициентом корреляции Пирсона.

Критерии согласия: хи-квадрат Пирсона, Колмогорова-Смирнова, Лиллиефорса. Применение критерия Шапиро-Уилки для проверки гипотезы о нормальности распределения.

• Если выбранные переменные имеют нормальный закон распределения, то в стартовой панели модуля **Основные статистики (Basic Statistics/ Tables)** выберите пункт **Correlation matrices (Корреляционные матрицы)**. Дважды щелкните по ней, либо

высветите и нажмите кнопку **OK**. На экране появится окно **Pearson Product-Moment Correlation (Корреляция Пирсона)** (рис. 2).

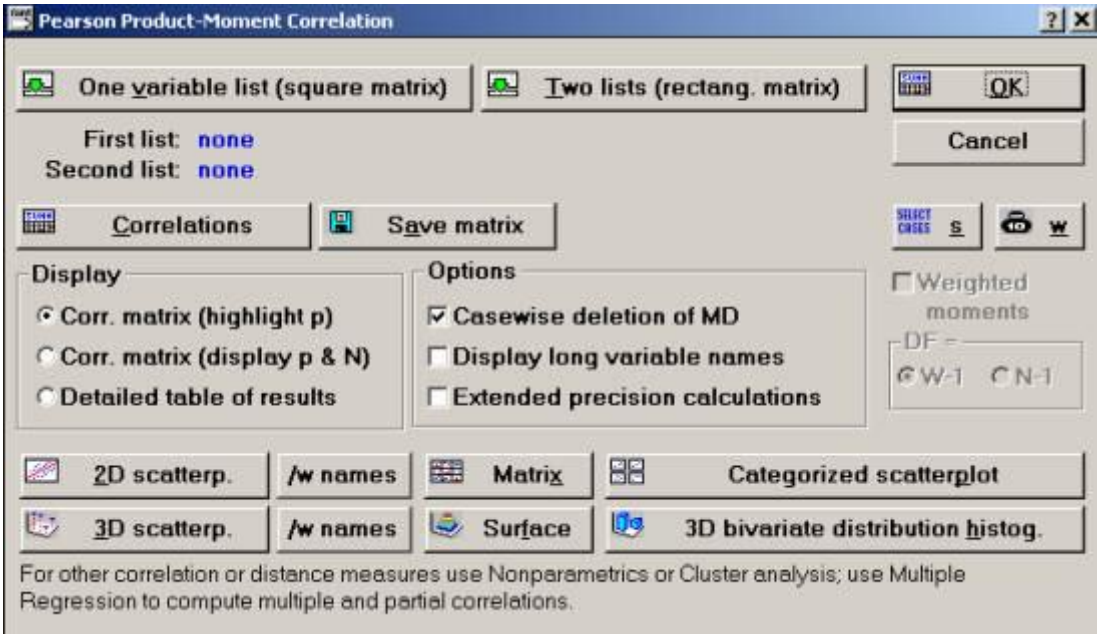


Рис. 2.

• Нажмите на кнопку **Two lists (Два списка)**. После чего откроется окно выбора переменных. Выберите переменные как показано на рисунке ниже (рис. 1.3). Таким образом, мы определили два списка переменных *X*4 –*X*6 – **First variables list (Первый список переменных)** и *Y*1 – **Second variables list (Второй список переменных)**. Мы хотим подсчитать корреляции между переменной *Y*1 и переменными *X*4 – *X*6. Щелкните по кнопке **OK** для подтверждения вашего выбора и возврата к окну **Pearson Product-Moment Correlation**.

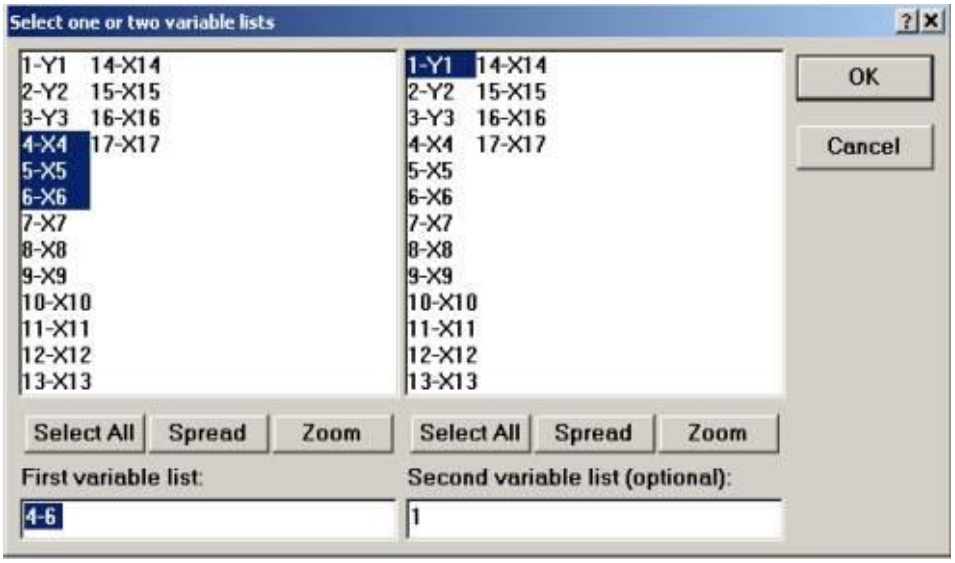


Рис. 3

- В окне **Pearson Product-Moment Correlation** нажмите кнопку **OK (Summary)**. На экране вы увидите корреляционную матрицу (рис. 1.4).

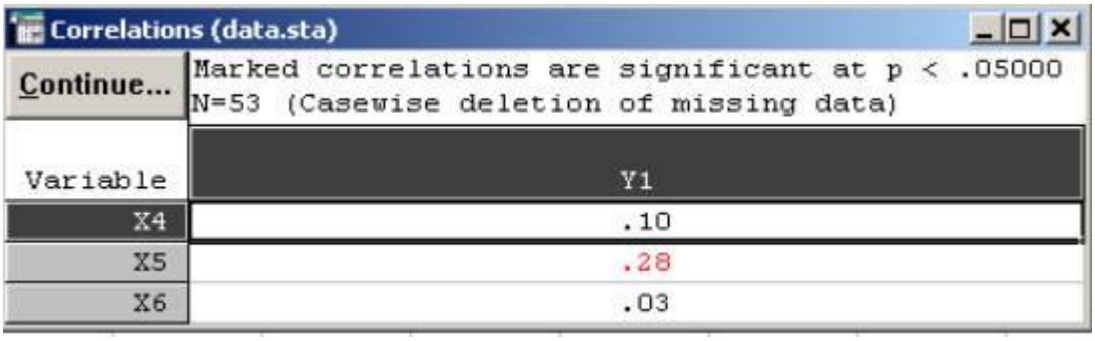


Рис. 4

В этой матрице имеется только один столбец, так как во втором списке мы выбрали только одну переменную. В столбце даны коэффициенты корреляции между переменной *Y*1 и *X*4 – *X*6. В нашей корреляционной матрице красным цветом автоматически выделены коэффициенты для уровня $p<0,05$. Именно на эти коэффициенты следует обратить наибольшее внимание. Грубо говоря, зависимость между переменными с выделенными красным цветом коэффициентами корреляции наиболее значимая. В нашем случае переменная *Y*1 наиболее зависима от переменной *X*5. Коэффициент

корреляции между этими переменными равен 0,28. Так как 0,28>0, то мы можем считать, что при возрастании переменной *X*5 переменная *Y*1 также возрастает. Рассмотрим эти переменные более внимательно. Полезно просмотреть зависимость между переменными *Y*1 и *X*5 графически.

- В окне корреляционной матрицы нажмите на кнопку **Continue (Продолжать)**. После чего вы вернетесь в окно **Pearson Product-Moment Correlation**. Нажмите на кнопку **Two lists (Два списка)**. После чего откроется окно выбора переменных (в данном случае для графика). Выберите переменные *X*5 и *Y*1 (рис. 5) и нажмите кнопку **OK**.

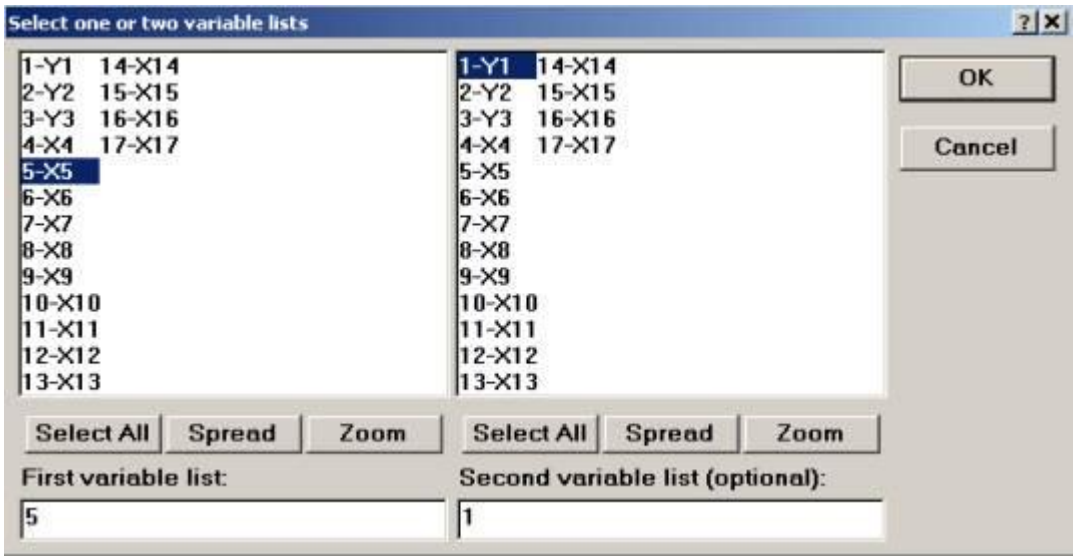


Рис. 5

- В окне **Pearson Product-Moment Correlation** нажмите кнопку **2D scatterplot (2D диаграмма рассеяния)**. После этого появится окно диаграммы рассеяния (рис. 6) для выбранных переменных.

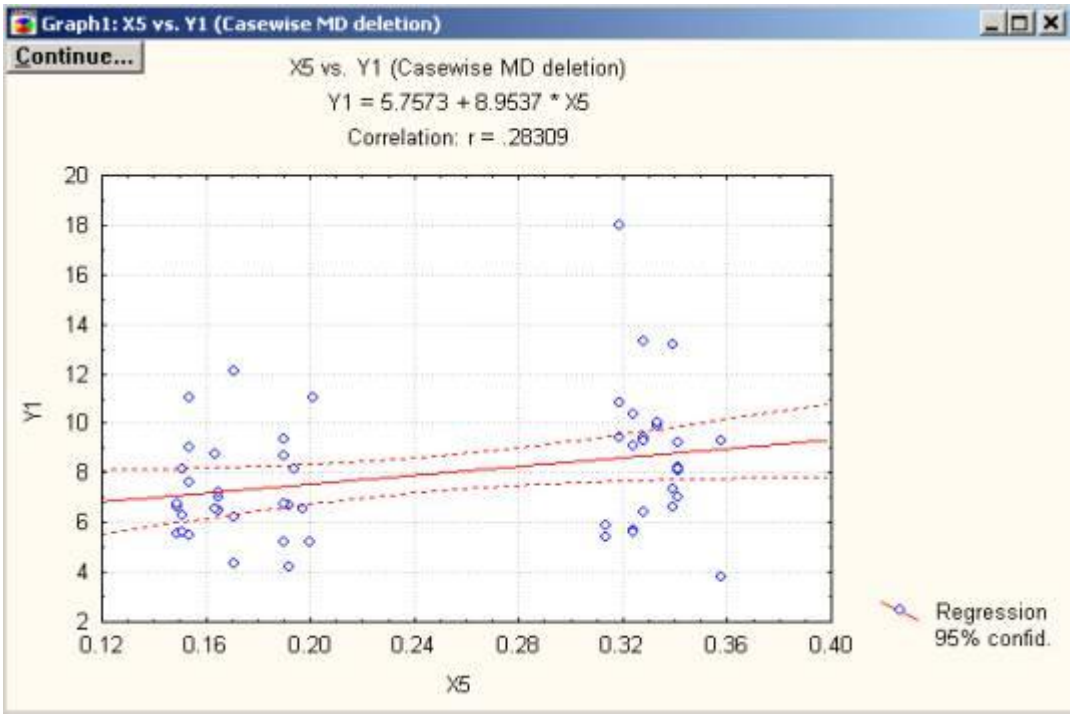


Рис. 6

- Из графика (рис. 6) отчетливо видно, что зависимость не является линейной – прямая очень плохо "ложится" на данные. На графике дана “лучшая” прямая. Как бы мы не меняли коэффициент наклона, подгонка будет только хуже. **STATISTICA** предлагает возможности, которые позволяют провести углубленное рассмотрение данных.

- Выберите средство **Кисть**, щелкнув по кнопке на инструментальной панели вверху. Перед вами справа появится панель **Brushing (Кисть)**. На панели **Brushing** сделайте установки как показано на рис. 7.

Рис. 7

• Войдите в график (просто щелкните по любой точке в его пространстве, сделав тем самым график активным) и отметьте лассо точки, которые, с вашей точки зрения, наиболее сильно отклоняются от прямой на графике. Мы выделяем точки с помощью лассо (обводя их карандашом, как бы захватывая лассо). Ранее была отмечена опция Lasso (Лассо) на панели инструментов Кисть. Выбрав, например, опцию Point (Точка), мы удаляли бы точки последовательно одна за другой. Выберите, например, точки над прямой, как показано на рис. 8.

• Щелкните далее на кнопку **Update (Обновить)** на панели **Brushing**, вы увидите следующий график (рис. 9).

Теперь данные лучше ложатся на прямую. Вы можете продолжить исследование. Вполне может оказаться, что в исключительных случаях имеется некоторая закономерность. Безусловно, эти закономерности стоит исследовать дополнительно. Вы легко можете определить, какие случаи были удалены вами. Для этого можно воспользоваться кнопкой **Label (Метка).**

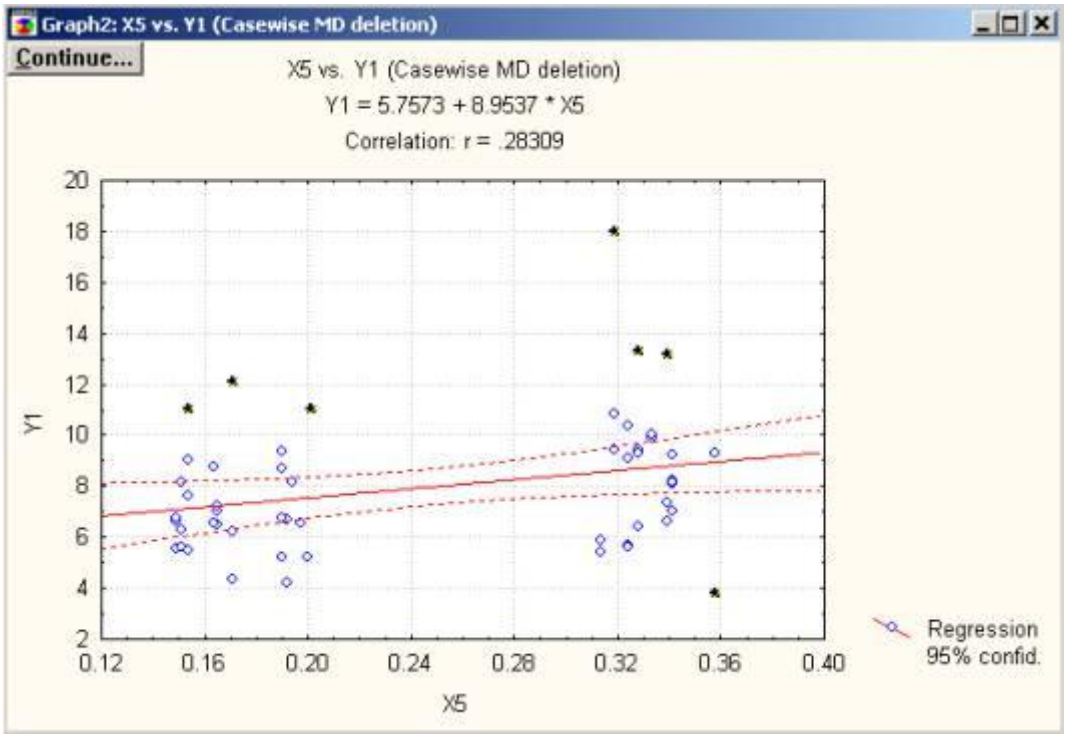

Graph2: X5 vs. Y1 (Casewise MD deletion)
Continue...
X5 vs. Y1 (Casewise MD deletion)
Y1 = 5.7573 + 8.9537 * X5
Correlation: r = .28309
Y1
X5
Regression
95% confid.

Рис. 8

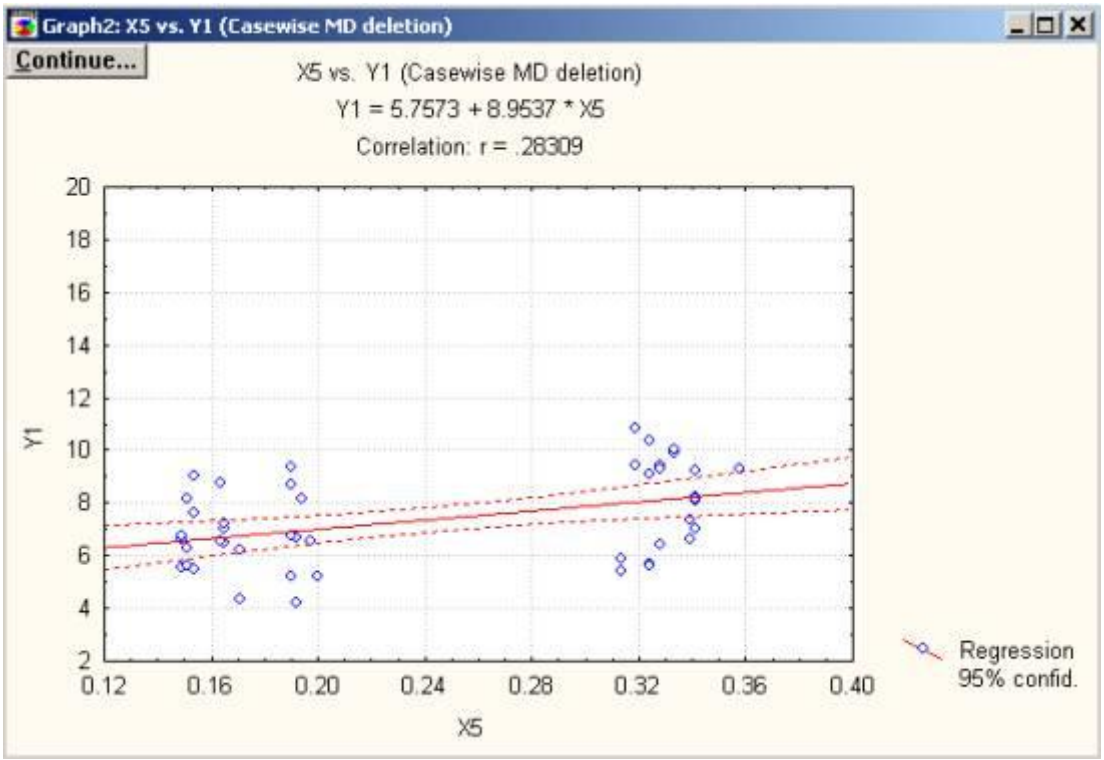

Graph2: X5 vs. Y1 (Casewise MD deletion)
Continue...
X5 vs. Y1 (Casewise MD deletion)
Y1 = 5.7573 + 8.9537 * X5
Correlation: r = .28309
Y1
X5
Regression
95% confid.

Рис. 9

• Щелкните на кнопку **De-select All (Reset All) (Отменить выбор всех)** вверху панели Brush. Пометьте опцию **Label** на панели **Brush** и вновь захватите лассо нужные точки. Далее нажмите кнопку **Update** и вы увидите на экране график, в котором рядом с выделенными точками появились имена случаев, к которым они относятся (рис. 10).

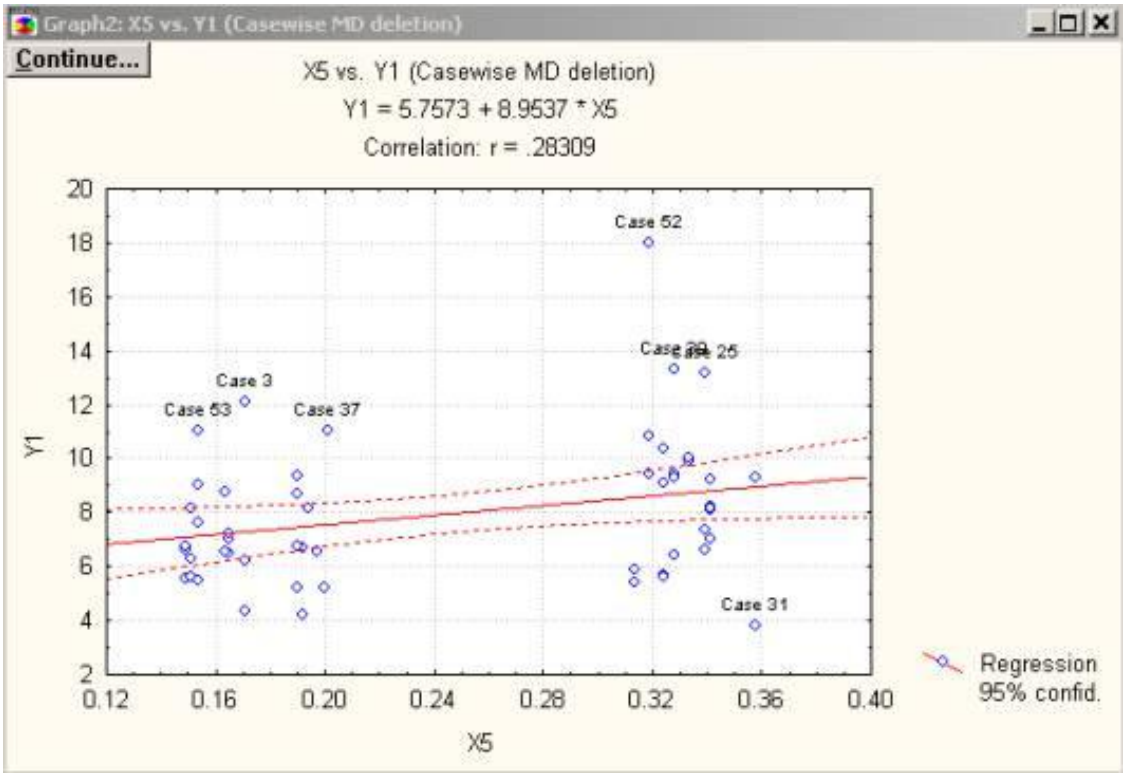


Рис. 10

Эти случаи как раз и требуют дополнительного исследования. Например, исключение их из рассмотрения может привести к значительному изменению исследуемого коэффициента корреляции. В том случае, если в корреляционной матрице имеются несколько высвеченных коэффициентов корреляции, то далее вам следует рассмотреть данные с другими высвеченными коэффициентами корреляции, построить графики зависимости, поработать с инструментом **Кисть**. Коэффициенты корреляции хорошо подходят для описания линейных связей и плохо, если зависимость между переменными не линейная. Вы можете просмотреть корреляционную

матрицу "графически" с помощью кнопки **Matrix (Матричный график)** в окне **Pearson Product-Moment Correlation**.

• Вернитесь в окно **Pearson Product-Moment Correlation** нажатием на кнопку **Continue**, расположенную на панели графика.

• Получите новую корреляционную матрицу для данных без лассо-точек. Проанализируйте изменение значений полученных коэффициентов корреляции.

•Вернитесь в окно **Pearson Product-Moment Correlation.**

• Щелкните по кнопке **One variable list (Один список переменных)**. Выберите переменные *Y*1, *X*4–*X*6 как показано на рис. 11, а затем нажмите кнопку **OK.**

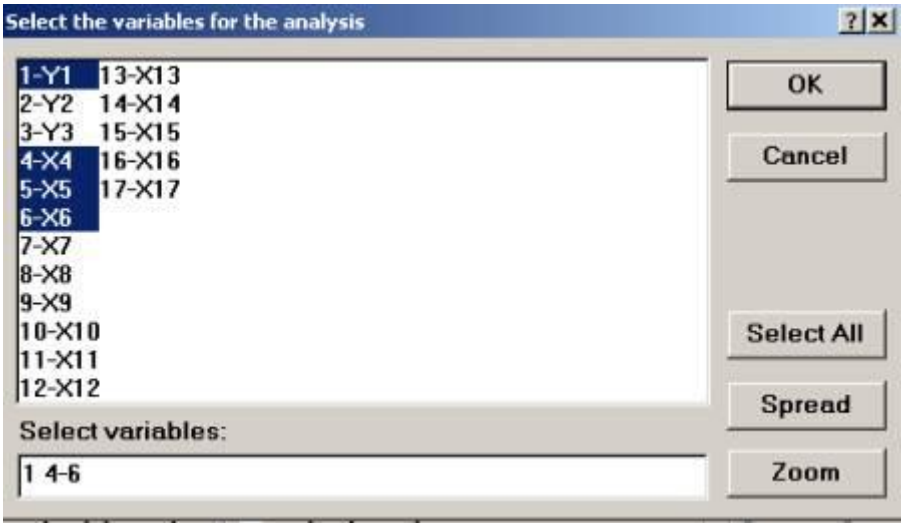


Рис. 11

• Далее в окне **Pearson Product-Moment Correlation** нажмите кнопку **Matrix**. После этого откроется окно выбора переменных для построения графиков (рис. 12).

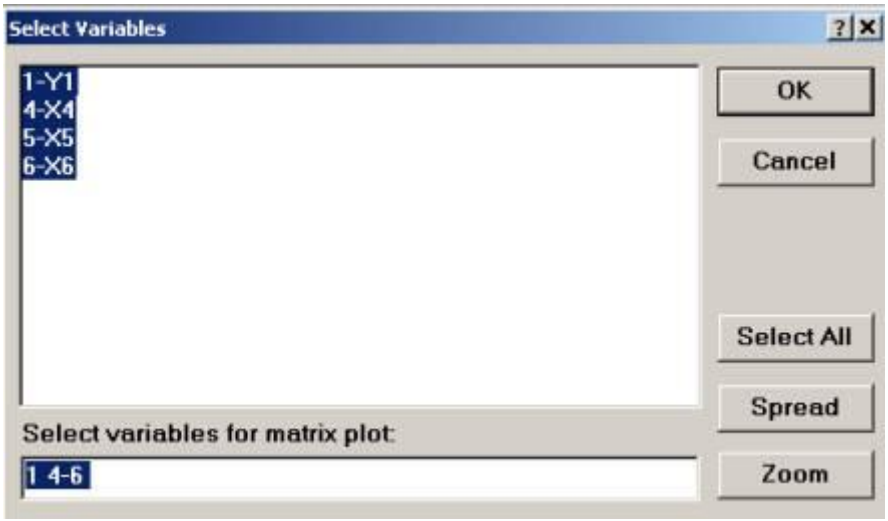


Рис. 12

• В окне выбора переменных выберите все переменные нажатием на кнопку **Select All (Выбрать все)**. Подтвердите свой выбор, нажав кнопку **OK**. На экране появится корреляционная матрица в графическом виде, позволяющая оценить линейные связи визуально (рис. 13).

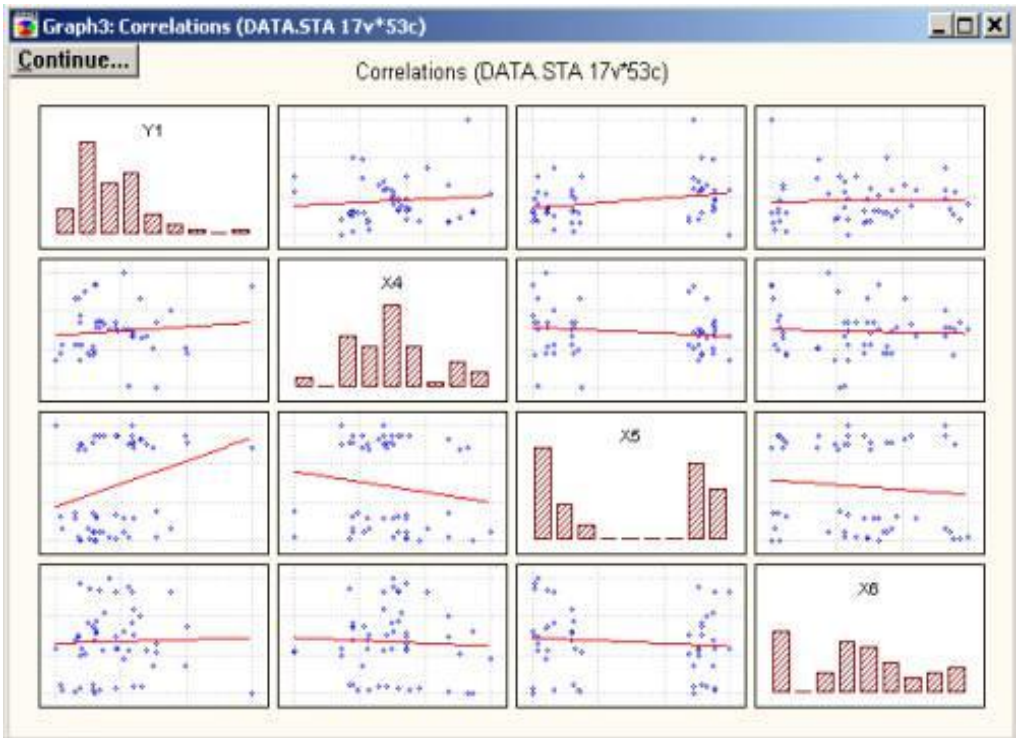


Рис. 13

2. Индивидуальные задания

• Для переменных, соответствующих вашему варианту (табл. 4), постройте корреляционную матрицу между всеми переменными (матрица размером 4x4). Определите максимальную корреляционную связь между переменными. Постройте корреляционную матрицу между всеми переменными "графически". Проанализируйте корреляционные связи между всеми переменными, с учетом знака и модуля связи (отрицательная или положительная связь, а также слабая, средняя и сильная связь). Определите значимость коэффициентов корреляции (Приложение);

• Для максимальной корреляции между двумя переменными (по модулю) постройте диаграмму рассеивания между двумя переменными.

• Определите на диаграмме точки, которые максимально удалены от линейной зависимости (не более 5). Удалите данные случаи из процесса корреляционного анализа.

• Пересчитайте коэффициент корреляции. Постройте новую диаграмму рассеивания.

• Сделайте вывод о том, как изменилась матрица после удаления данных.

Таблица 4

Номер варианта	Переменные	Номер варианта	Переменные
1	*Y*1,*Y*2,*X*4,*X*5	13	*Y*2,*Y*3,*X*14,*X*15
2	*Y*1,*Y*2,*X*6,*X*7	14	*Y*2,*Y*3,*X*16,*X*17
3	*Y*1,*Y*2,*X*8,*X*9	15	*Y*1,*Y*3,*X*4,*X*5
4	*Y*1,*Y*2,*X*10,*X*11	16	*Y*1,*X*4,*X*15,*X*3
5	*Y*1,*Y*2,*X*12,*X*13	17	*Y*1,*Y*2,*X*8,*X*15
6	*Y*1,*Y*2,*X*14,*X*15	18	*Y*2,*Y*3,*X*6,*X*11
7	*Y*1,*Y*2,*X*16,*X*17	19	*Y*1,*Y*3,*X*14,*X*17
8	*Y*2,*Y*3,*X*4,*X*5	20	*Y*2,*Y*3,*X*7,*X*10
9	*Y*2,*Y*3,*X*6,*X*7	21	*Y*1,*Y*2,*X*6,*X*12
10	*Y*2,*Y*3,*X*8,*X*9	22	*Y*1,*Y*3,*X*6,*X*12
11	*Y*2,*Y*3,*X*10,*X*11	23	*Y*2,*Y*3,*X*7,*X*12
12	*Y*2,*Y*3,*X*12,*X*13	24	*Y*1,*Y*3,*X*7,*X*16

ПРИЛОЖЕНИЕ

Критические значения коэффициента корреляции

Число степеней свободы, f	Критическое значение, R_{KP}	Число степеней свободы, f	Критическое значение, R_{KP}	Число степеней свободы, f	Критическое значение, R_{KP}
1	0,997	9	0,602	17	0,456
2	0,950	10	0,576	18	0,444
3	0,878	11	0,553	19	0,433
4	0,811	12	0,532	20	0,423
5	0,754	13	0,514	30	0,349
6	0,707	14	0,497	50	0,273
7	0,666	15	0,482	80	0,217
8	0,632	16	0,468	100	0,195

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Айвазян С.А. Прикладная статистика. Основы моделирования и первичная обработка данных / С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин. М.: Финансы и статистика,1983.

2. Коровин Е.Н. Методы обработки биомедицинских данных: учеб. пособие / Е.Н. Коровин, О.В. Родионов. Воронеж: ГОУВПО «Воронежский государственный технический университет», 2007.

3. Фестер Э. Методы корреляционного и регрессионного анализа / Э. Фестер, В. Ренц. М.: Финансы и статистика, 1983.

ЛАБОРАТОРНАЯ РАБОТА № 2

Регрессионный анализ

Цель работы: Построить регрессионную модель и провести полный регрессионный анализ.

1. Общие сведения. Простая линейная регрессия

Функция регрессии

В регрессионном анализе изучается связь и определяется количественная зависимость между зависимой переменной и одной или несколькими независимыми переменными. Пусть переменная *Y* зависит от одной переменной x. При этом предполагается, что переменная x принимает заданные фиксированные значения, а зависимая переменна *Y* имеет случайный разброс из-за ошибок измерения, влияния неучтенных факторов и т.д. Каждому значению x соответствует некоторый закон распределения вероятностей случайной величины *Y*. Предположим, что *Y* в "среднем" линейно зависит от значений переменной x. Это означает, что условное математическое ожидание случайной величины Y при заданном значении x имеет вид

$$M(Y/x) = a_0 + a_1 x \qquad (1)$$

Данная функция называется линейной теоретической функцией регрессии Y на x, а параметры a_0 и a_1 – параметрами линейной регрессии (коэффициенты регрессии). На практике параметры регрессии определяются по результатам наблюдений переменных Y и x, связь между которыми можно записать

$$Y = a_0 + a_1 x + \varepsilon \quad , \tag{2}$$

где ε – случайная ошибка наблюдений.

Последовательность проведения регрессионного анализа

- Формулировка задачи.
- Идентификация переменных (определение входных и выходных переменных).
- Сбор статистических данных.
- Спецификация функции регрессии (определение вида модели).
- Оценивание параметров функции регрессии.
- Оценка точности регрессионного анализа:

1) Проверка адекватности всей модели, т.е. согласуются ли предсказанные значения выходной величины с наблюдаемыми данными;

2) Проверка значимости параметров модели, т.е. значимо ли они отличаются от нуля или нет.

- Интерполяция результатов, анализ, оптимизация и прогнозирование.

Предпосылки к проведению регрессионного анализа

- Случайные ошибки наблюдений имеют нормальный закон распределения

$$\varepsilon \to N(0,\sigma), \quad M(\varepsilon) = 0, \quad D(\varepsilon) = \sigma^2 = const. \tag{3}$$

- Отсутствие автокорреляции между ошибками наблюдений, т.е. последовательные значения ε_i не зависят друг от друга.

Метод наименьших квадратов

Для нахождения оценок параметров модели по результатам наблюдений используется метод наименьших квадратов (МНК). Пусть проведено *n* независимых наблюдений случайной величины *Y* при соответствующих значениях x, совместный закон распределения которых неизвестен. Следовательно, теоретическую функцию регрессии мы не сможем найти. Наша задача оценить эмпирическую функцию регрессии

$$\tilde{y} = \tilde{a}_0 + \tilde{a}_1 x. \qquad (4)$$

Согласно МНК, параметры подбираются таким образом, чтобы минимизировать сумму квадратов отклонений наблюдаемых значений от расчетных по модели значений

$$F = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 = \sum_{i=1}^{n} (y_i - \tilde{a}_0 - \tilde{a}_1 x_i)^2 \to \min,$$

где y_i – наблюдаемые значения выходной переменной; $\tilde{y}_i$ – значения выходной переменной, рассчитанные по модели.

Из необходимых условий минимума

$$\begin{cases} \dfrac{\partial F}{\partial \tilde{a}_0} = -2 \sum\limits_{i=1}^{n} (y_i - \tilde{a}_0 - \tilde{a}_1 x_i) = 0, \\ \dfrac{\partial F}{\partial \tilde{a}_1} = -2 \sum\limits_{i=1}^{n} (y_i - \tilde{a}_0 - \tilde{a}_1 x_i) x_i = 0 \end{cases} \qquad (5)$$

находим оценки параметров a_0 и a_1 (здесь и далее, если это не мешает пониманию, знак ~ над параметрами будет опускаться). Они будут определяться из решения системы двух линейных уравнений

$$\begin{cases} n a_0 + a_1 \sum x_i = \sum y_i, \\ a_0 \sum x_i + a_1 \sum x_i^2 = \sum y_i x_i. \end{cases} \quad (6)$$

Здесь и далее, если это не оговорено особо, суммирование происходит от i=1,n. Оценки параметров, получаемые по методу МНК, при условии выполнения предпосылок относительно случайных ошибок наблюдений, будут обладать следующими свойствами:

- несмещенность;
- состоятельность;
- эффективность.

Проверка адекватности модели

Для проверки гипотезы адекватности модели необходимо сравнить две суммы квадратов:

1) Остаточную сумму квадратов, характеризующую отклонение от регрессии

$$Q_e = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2. \quad (7)$$

2) Сумму квадратов, обусловленную регрессией

$$Q_R = \sum_{i=1}^{n} (\tilde{y}_i - \bar{y})^2, \quad (8)$$

где $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$.

Тогда выборочное значение ***F,*** имеющее распределение Фишера

$$F = \frac{Q_R / k}{Q_e / (n-k-1)}, \tag{9}$$

может служить проверкой адекватности для заданного уровня значимости λ (обычно для экономических задач λ=0,05) и степеней свободы $f_1 = k$; $f_2 = n-k-1$, где k – число оцениваемых параметров, исключая свободный коэффициент.

Если $F \geq F_{\lambda;\, f1;\, f2}$ – модель адекватна (прил.1). Остаточную дисперсию ошибки

$$S^2 = Q_e / (n-k-1) \tag{10}$$

можно использовать в качестве оценки дисперсии σ^2 – дисперсии случайной величины. Результаты проверки адекватности удобно представить в виде таблицы (табл.1).

Полезной характеристикой линейной регрессии является коэффициент детерминации, вычисляемый по формуле

$$R^2 = \frac{Q_R}{Q_R + Q_e} = 1 - \frac{Q_e}{Q_R + Q_e}. \tag{11}$$

Таблица 1

Источник изменения	Сумма квадратов	Число степеней свободы	Оценка дисперсии
Модель	$Q_R = \sum_{i=1}^{n} (\bar{y}_i - \bar{y})^2$	k	$S_R^2 = Q_R / k$
Ошибка	$Q_e = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$	$n-k-1$	$S^2 = Q_e/(n-k-1)$
Сумма	$Q_e + Q_R = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$n-1$	

Коэффициент детерминации равен той доле результатов наблюдений относительно горизонтальной прямой $y = \bar{y}$, которая объясняется уравнением регрессии. Величина $R = +\sqrt{R^2}$ является оценкой множественного коэффициента корреляции между результатами наблюдений и вычисленными значениями $\hat{y}_i$. Если R^2=0.75 это значит, что модель работает на 75%, а 25% приходится на ошибку или неучтенные в модели факторы (для практических целей целесообразно, чтобы $R^2 \geq 0{,}75$). Для небольших значений n<30 необходимо использовать скорректированный коэффициент детерминации

$$R^{*2} = 1 - \frac{n-1}{n-k-1}(1-R^2). \qquad (12)$$

Проверка значимости параметров модели

В результате проверки устанавливается статистическая значимость или незначимость отличия от нуля оценок параметров регрессии. Это проверка осуществляется отдельно для каждого параметра модели. Для оценки значимости коэффициентов регрессии можно воспользоваться следующим правилом, если абсолютная величина коэффициента регрессии больше доверительного интервала, то гипотеза о незначимости коэффициента отвергается

$$\left|\tilde{a}_i\right| \geq t_{f,\,\lambda/2}\, S_{a_i}\ , \qquad (13)$$

где $t_{f,\lambda/2}$ – значение Стьюдента, определяемое по числу степеней свободы $f = n-k-1$ и λ=0,05 (прил. 2); S_{a_i} – среднеквадратичные отклонения ошибок коэффициентов регрессии.

Для простой линейной регрессии $y = a_0 + a_1 x$ они могут быть вычислены соответственно

$$S_{a_0} = \sqrt{\frac{S^2 \sum x^2}{n\sum x^2 - (\sum x)^2}}, \qquad S_{a_1} = \sqrt{\frac{nS^2}{n\sum x^2 - (\sum x)^2}}. \quad (14)$$

Можно проверять значимость коэффициентов по *t*-критерию Стьюдента:

$$t = \frac{|\tilde{a}_i|}{S_{a_i}}. \quad (15)$$

Вычисленное значение сравнивается с табличным (прил. 2) и если $t \geq t_{f;\ \lambda/2,}$ то коэффициент значим. В противном случае соответствующую переменную можно исключить из модели и все расчеты, включая решение системы линейных уравнений, повторить снова.

Множественная линейная регрессия представляет собой выражение:

$$\tilde{y} = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_k x_k. \quad (16)$$

Некоторые нелинейные модели, сводящиеся к линейным

Существуют два вида нелинейности регрессионных моделей:

1) Нелинейные относительно независимых переменных.

Например,

$\tilde{y} = a_0 + a_1 x_1 + a_2 x_2^2 + a_3 x_1 x_2$.

В этом случае необходимо просто сделать замену переменных:

$x_2^2 = z_1$, $x_1 x_2 = z_2$,

$\tilde{y} = a_0 + a_1 x_1 + a_2 z_1 + a_3 z_2$.

2) Нелинейные относительно параметров регрессии.

Например,

$$y = \frac{1}{a_0 + a_1 x}.$$

Выполним функциональное преобразование:

пусть $z = \frac{1}{y}$, тогда $z = a_0 + a_1 x.$

К сожалению, не всегда можно функциональными преобразованиями от нелинейных моделей перейти к линейным. Кроме того, нужно иметь в виду, что при вычислении параметров по методу МНК минимизируется сумма квадратов отклонений преобразованных, а не исходных данных.

Проверка предпосылок регрессионного анализа

Проверка нормальности закона распределения ошибок

Анализ ошибок проводится по следующей схеме. Предполагаем, что $\varepsilon_i \sim N(0,\sigma)$, тогда $\varepsilon_i \sim N(0,1)$. Тогда, если модель правильна, то дисперсия остатков, характеризующая качество аппроксимации результатов наблюдений

$$S^2 = \frac{Q_e}{n-k-1} = \frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-k-1} = \frac{\sum_{i=1}^{n} e_i^2}{n-k-1}, \qquad (17)$$

служит оценкой величины σ^2 – дисперсии ошибок наблюдений, где $\bar{e}$ – среднее значение отклонений. Случайная величина e_i / S

представляет собой единичные нормальные отклонения. Если эти отклонения будут находиться в интервале [-2; 2] , то, следовательно, наше предположение о том, что $\varepsilon_i \sim N(0,\sigma)$ не ошибочно.

Проверка на однородность случайных ошибок

Для проверки на однородность дисперсии $D(\varepsilon)=const$ целесообразно воспользоваться методом Гольфельда. Последовательность значений случайной величины Y разбивается на две последовательности объемом n_1 и n_2 ($n_1+n_2=n$). Для каждой последовательности вычисляются дисперсии воспроизводимости S_1^2 и S_2^2. Тогда отношение

$$F = \frac{S_2^2}{S_1^2} \qquad (18)$$

при $S_1^2 < S_2^2$ будет иметь распределение Фишера со степенями свободы $f_1=n_1-k-1$, $f_2=n_2-k-1$. Если значение F превышает табличное, то гипотеза об однородности дисперсии отклоняется. Чувствительность критерия увеличивается, если исключить средние наблюдения.

В случае если дисперсии оказались неоднородными, часто оказывается полезным изменение масштаба для выходной переменной. Вводится некоторая функция от выходной переменной, например $\ln y$ или $\sqrt{y}$.

Проверка на автокорреляцию случайных ошибок

Наличие автокорреляции ошибок можно проверить с помощью критерия Дарбина-Уотсона:

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}. \qquad (19)$$

Критерий Дарбина-Уотсона изменяется в диапазоне $0 \le DW \le 4$. При отсутствии автокорреляции DW=2. В приложении 4 приводятся нижние и верхние границы критерия dн и dв для степеней свободы $f_1=n$; $f_2=k$.

Если:

$0 \le DW \le d$н, есть положительная автокорреляция,

$4-d$н $\le DW \le 4$, есть отрицательная автокорреляция,

dв $\le$ DW $\le 4-d$в, автокорреляция отсутствует,

dн$< DW < d$в или $4-d$в $\le DW \le 4-d$н, нужны дополнительные исследования.

2. Описание типового примера

В табл. 2 приведены значения выходной переменной y_i при данном значении входной переменной x_i. Модель ищется в виде

$$\bar{y} = a_0 + a_1 x.$$

Таблица 2

X	1	2	3	4
Y	2	4	5	7

1. Вычислим оценки математических ожиданий, смещенные и несмещенные оценки дисперсий и средних квадратичных отклонений:

$$\bar{x} = \frac{1}{n}\sum x_i = 2{,}5\ , \quad \bar{y} = \frac{1}{n}\sum y_i = 4{,}5\ ,$$

$$S_x^2 = \frac{1}{n}\sum (x_i - \bar{x})^2 = 1{,}250\ , \quad S_y^2 = \frac{1}{n}\sum (y_i - \bar{y})^2 = 3{,}251,$$

$$S_{xн}^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2 = 1{,}666\ , \quad S_{yн}^2 = \frac{1}{n}\sum (y_i - \bar{y})^2 = 4{,}333,$$

$$S_x = \sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2} = 1{,}118\ , \quad S_y = \sqrt{\frac{1}{n}\sum (y_i - \bar{y})^2} = 1{,}803,$$

$$S_{xн} = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2} = 1{,}291\ , \quad S_{yн} = \sqrt{\frac{1}{n-1}\sum (y_i - \bar{y})^2} = 2{,}082.$$

2. Определим модель. Вычислим соответствующие суммы и составим систему двух линейных уравнений:

$$\Sigma x_i = 10, \quad \Sigma y_i = 18, \quad \Sigma x_i^2 = 30, \quad \Sigma x_i y_i = 53, \quad \Sigma y_i^2 = 94,$$

$$\begin{cases} 4a_0 + 10a_1 = 18, \\ 10a_0 + 30a_1 = 53, \end{cases}'$$

решая которую, получаем: $a_0 = 0{,}5$, $a_1 = 1{,}6$. Модель имеет вид

$$\hat{y} = 0{,}5 + 1{,}6x$$

3.Рассчитаем модельные значения, подставляя в уравнение значения входной переменной (табл. 3).

Таблица 3

$\hat{y}$	2,1	3,7	5,3	6,9

4. Определим адекватность модели, для этого вычислим общую сумму квадратов, сумму квадратов, относящуюся к регрессии, и сумму квадратов остатков:

$$Q = (y_i - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n} = 94 - \frac{(18)^2}{4} = 13;$$

$$Q_R = \sum (\hat{y}_i - \bar{y})^2 = \sum \hat{y}^2 - \frac{(\sum y)^2}{n} = 93{,}8 - \frac{(18)^2}{4} = 12{,}8;$$

$$Q_e = \Sigma e_i^2 = (2-2{,}1)^2 + (4-3{,}7)^2 + (5-5{,}3)^2 + (7-6{,}9)^2 = 0{,}02 .$$

Для проверки полученных расчетов необходимо, чтобы $Q = Q_R + Q_e$. В нашем случае равенство выполняется.

Расчетное значение критерия Фишера равно:

$$F = \frac{Q_R / k}{Q_e / (n-k-1)} = \frac{S_R^2}{S^2} = \frac{12{,}8/1}{0{,}2/(4-1-1)} = 128$$

Расчетное значение критерия Фишера больше, чем табличное значение $F > F_{\text{ТАБЛ.}} = F_{0{,}05;1;2} = 18{,}512$ (приложение 1).

Вывод 1: модель адекватна, исходные данные хорошо согласуются с моделью.

5. Оперативно адекватность модели можно проверить по коэффициентам детерминации или корреляции. Коэффициент детерминации, скорректированный коэффициент детерминации и коэффициент корреляции определяются следующим образом:

$$R^2 = \frac{Q_R}{Q_R + Q_e} = \frac{Q_R}{Q} = \frac{12{,}8}{13} = 0{,}9846,$$

$$R^{*2} = 1 - \frac{n-1}{n-k-1}(1-R^2) = 1 - \frac{4-1}{4-1-1}(1-0{,}9846) = 0{,}977,$$

$$R = \sqrt{R^2} = 0{,}9922.$$

Полученное значение коэффициента детерминации больше, чем 0,75 следовательно, модель можно считать адекватной.

8. Определим значимость параметров модели. Вычислим средние квадратические отклонения ошибок коэффициентов регрессии:

$$S_{a0}^2 = S^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right)^2 = \frac{S^2 \sum x_i^2}{n\sum x_i^2 - (\sum x_i)^2} = \frac{0{,}1 \cdot 30}{4 \cdot 30 - (10)^2} = 0{,}15;$$

$$S_{a1}^2 = \frac{s^2}{\sum(x_i - \bar{x})^2} = \frac{nS^2}{n\sum x_i^2 - (\sum x_i)^2} = \frac{4 \cdot 0{,}1}{4 \cdot 30 - (10)^2} = 0{,}02;$$

$$S_{a0} = 0{,}387 \quad ; \quad S_{a1} = 0{,}141 \ .$$

Табличное значение Стьюдента $t_{2;0,025}$=4,303 (приложение 2). Получаем расчетные значения Стьюдентов:

$$t_{a_0} = 0{,}5/0{,}387 = 1{,}291;$$

$$t_{a_1} = 1{,}6/0{,}141 = 11{,}31.$$

Вывод 2: параметр a_1 будет значим так, расчетное значение Стьюдента будет больше, чем табличное, параметр a_0 – незначим.

Аналогичный вывод можно сделать иначе, вычисляя доверительные интервалы для параметров модели:

$$a_0 = 0{,}5 \pm 4{,}303 \cdot 0{,}387 = 0{,}5 \pm 1{,}665;$$

$$a_1 = 1{,}6 \pm 4{,}303 \cdot 0{,}141 = 1{,}6 \pm 0{,}607.$$

Так как абсолютное значение параметра a_1 будет больше, чем его доверительный интервал, то параметр значим. Параметр a_0 незначим.

9.Оценка среднеквадратического отклонения ошибки $S = \sqrt{0{,}1} = 0{,}316$. Разделив все отклонения (остатки) на эту величину, получим нормированные отклонения, которые все находятся в интервале [-2; +2] (табл. 3).

Вывод 3: первая предпосылка выполняется, случайная ошибка имеет нормальный закон распределения с нулевым математическим ожиданием.

10. Критерий Дарбина-Уотсона вычисляем следующим образом:

$$DW = ((0{,}3+0{,}1)^2+(-0{,}3-0{,}3)^2+(0{,}1+0{,}3)^2) / 0{,}2=3{,}4 ;$$

4-dв < DW < 4 (dB примерно равно 1,5 (приложение 3)).

Вывод 4: вторая предпосылка не выполняется, между текущими значениями случайной величины присутствует отрицательная автокорреляция.

В завершении работы приведем итоговую таблицу остатков (табл. 4).

Таблица 4

N/N	Значение Y	Оценка Y	Остаток	Нормированные остатки
1	2	2,1	–0,1	–0,316
2	4	3,7	0,3	0,948
3	5	5,3	–0,3	–0,948
4	7	6,9	0,1	0,316

3. Решение типового примера с применением Statistica

Для проведения расчетов с модулем предварительной обработки необходимо выполнить следующие действия:

• запустить программу **STATISTICA** командой ***Пуск/Программы /Statistica/Statistica;***

• выделить строку **Basic Statistica/Tables** и нажать кнопку **Switch To**;

• в открывшемся окне **Basic Statistica/Tables** закрыть все окна документов и выполнить команду ***File/New Data***. Появится таблица для ввода данных (рис. 1);

• введите исходные данные для переменных X и Y в столбцы VAR1 и VAR2 соответственно варианту задания (табл. 5);

• выделите блок колонок VAR3-VAR10 и нажмите на панели инструментов кнопку **Vars**. В появившемся меню выберите команду ***Delete***, а затем в диалоговом окне нажмите кнопку **OK**;

• выделите блок строк 5–10 и нажмите на панели инструментов кнопку **Cases**. В появившемся меню выберите команду ***Delete***, а затем в диалоговом окне нажмите кнопку **OK**;

• щелкните правой клавишей мыши по столбцу VAR1 и выберите команду Variable Specs.... Появится диалоговое окно (рис. 2);

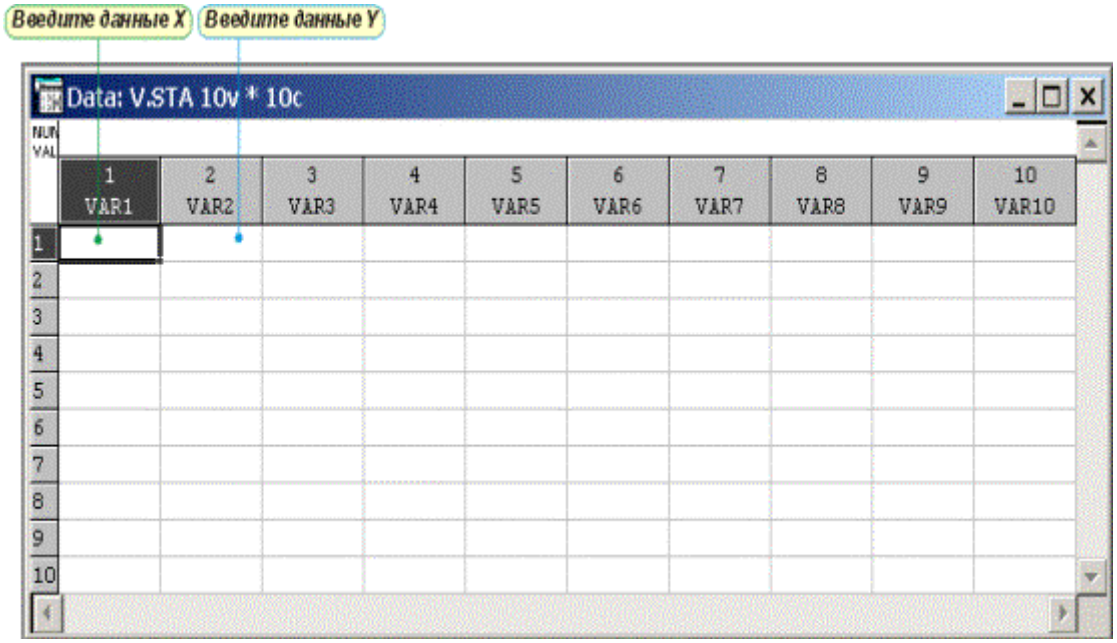


Рис. 1

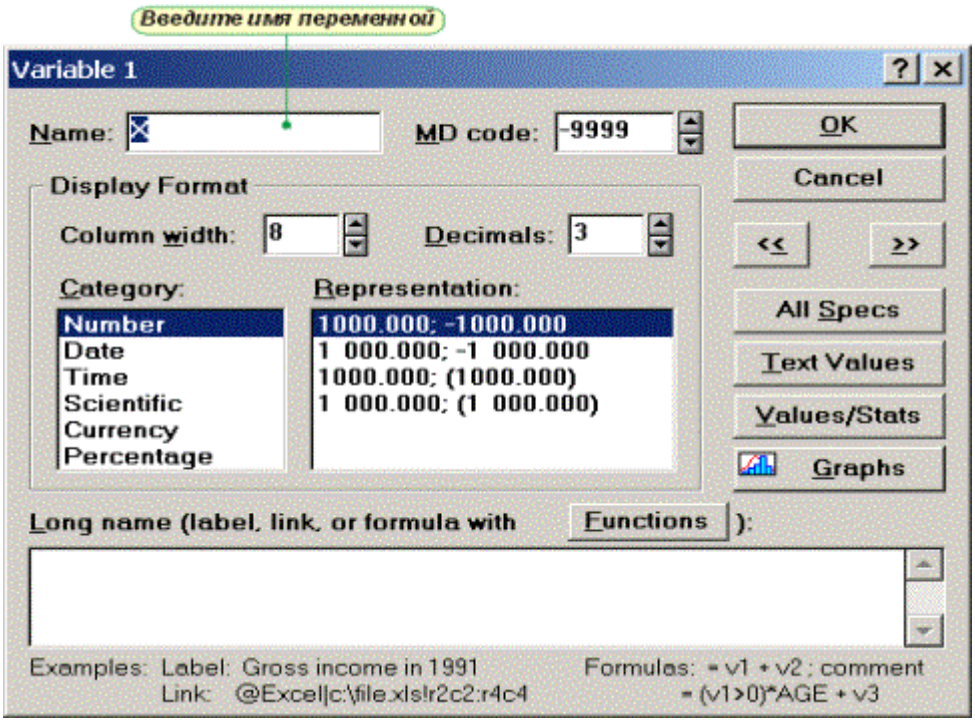


Рис. 2

- введите имя переменной *X* и нажмите **OK**;
- щелкните правой клавишей мыши по столбцу VAR2 и выберите команду ***Variable Specs....,*** введите имя переменной *Y* и нажмите **OK**;
- в результате вы должны получить таблицу данных следующего вида (рис. 3);

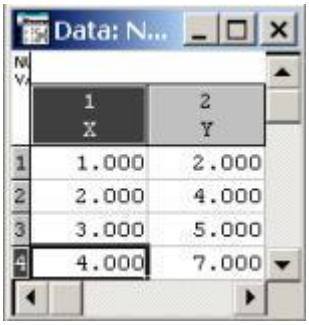


Рис. 3

- сохраните полученный файл данных командой ***File/Save As*** под именем *lab*1.*sta* в предназначенной для работы папке;

• сделайте активным окно с таблицей данных, а затем выполните команду ***Analysis/Resume Analysis***. Появится окно (рис. 4).

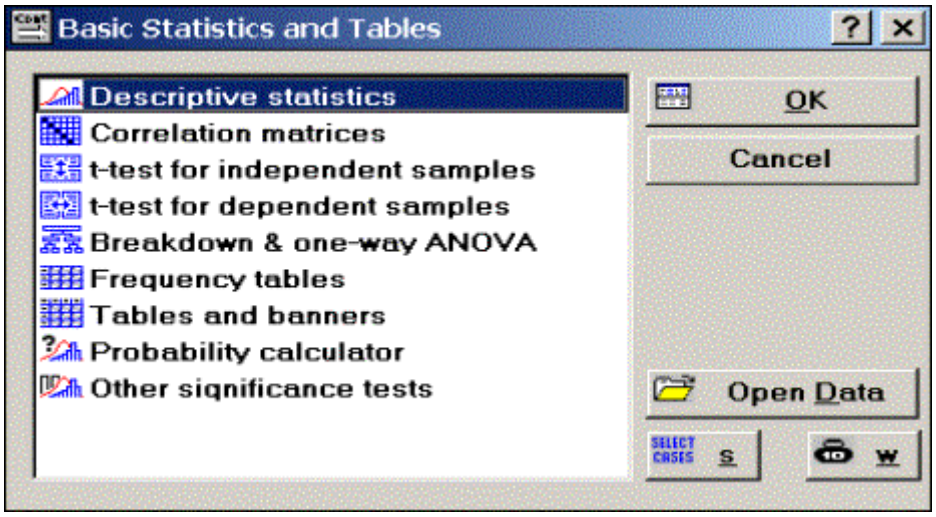


Рис. 4

• выберете пункт **Descriptive statistics** и нажмите **OK**. Появится окно (рис. 5);

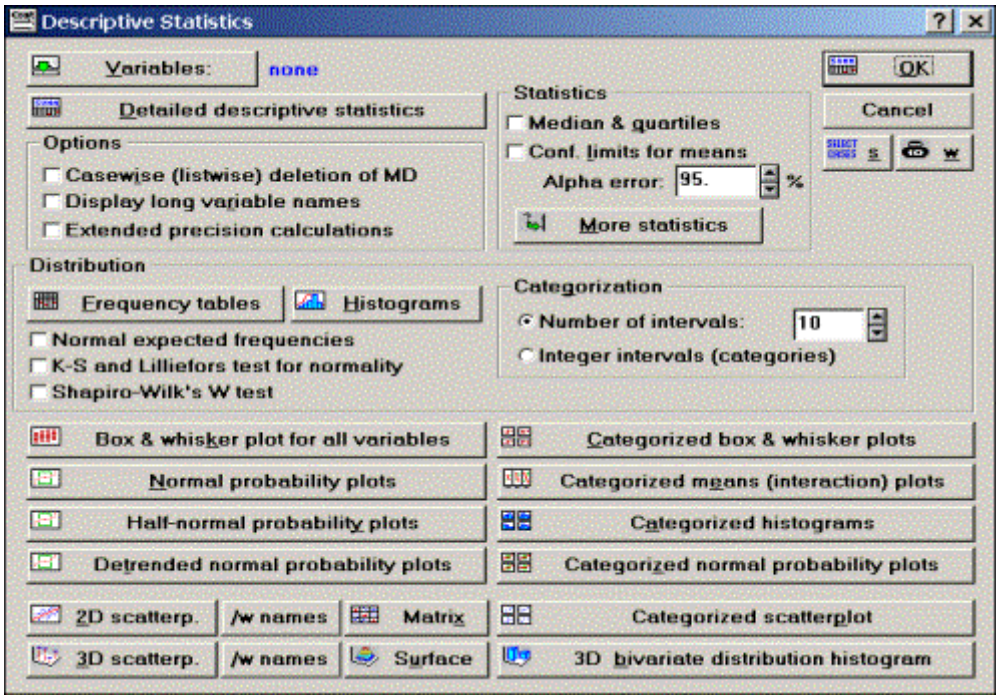


Рис. 5

• нажмите кнопку **Variables**. В появившемся диалоговом окне (рис. 6) нажмите кнопку **Select All** и **OK**;

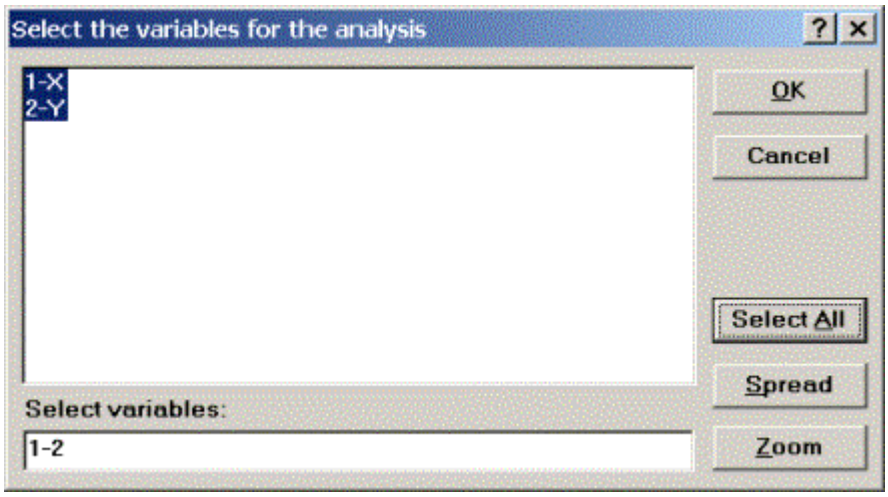


Рис. 6

• программа выполнит переход в предыдущее окно, в котором необходимо нажать кнопку **Detailed descriptive statistics**. Появится таблица с результатами расчетов (рис. 7);

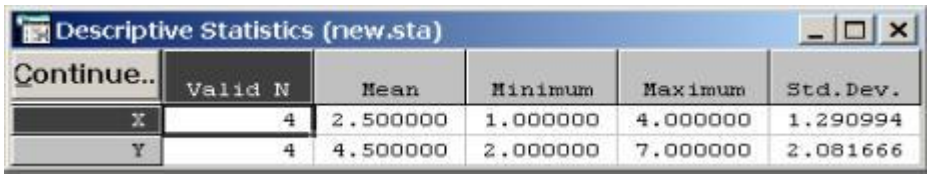


Рис. 7

• выпишите необходимые данные;

• нажмите кнопку **Continue**. Выполнится переход в предыдущее окно, нажмите кнопку **2D scatterp.** Появится окно для выбора переменных (рис. 8);

• сделайте необходимые изменения и нажмите кнопку **OK**;

• появится окно (рис. 9) с графиком линейной регрессионной модели, с рассчитанными значениями коэффициентов модели и коэффициентом корреляции. Первая часть работы выполнена.

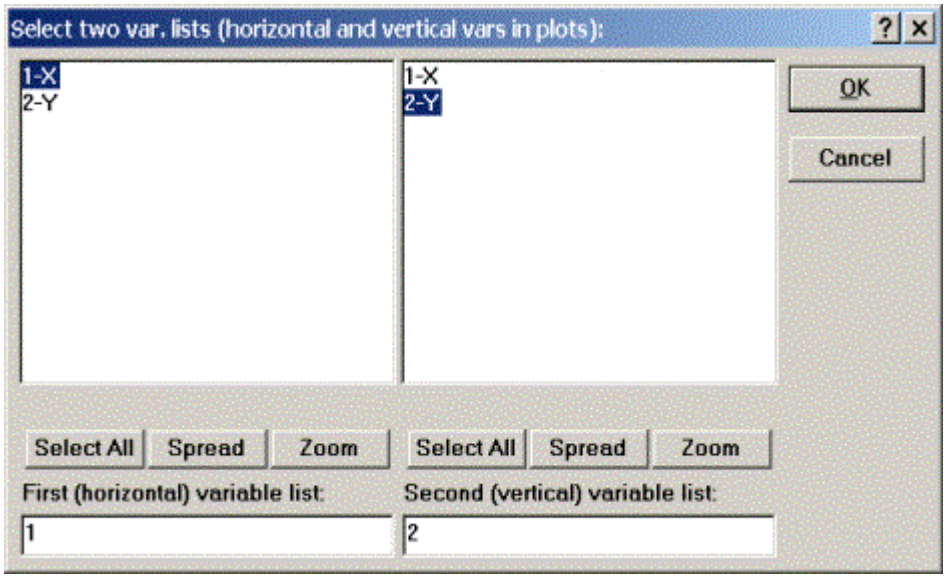


Рис. 8

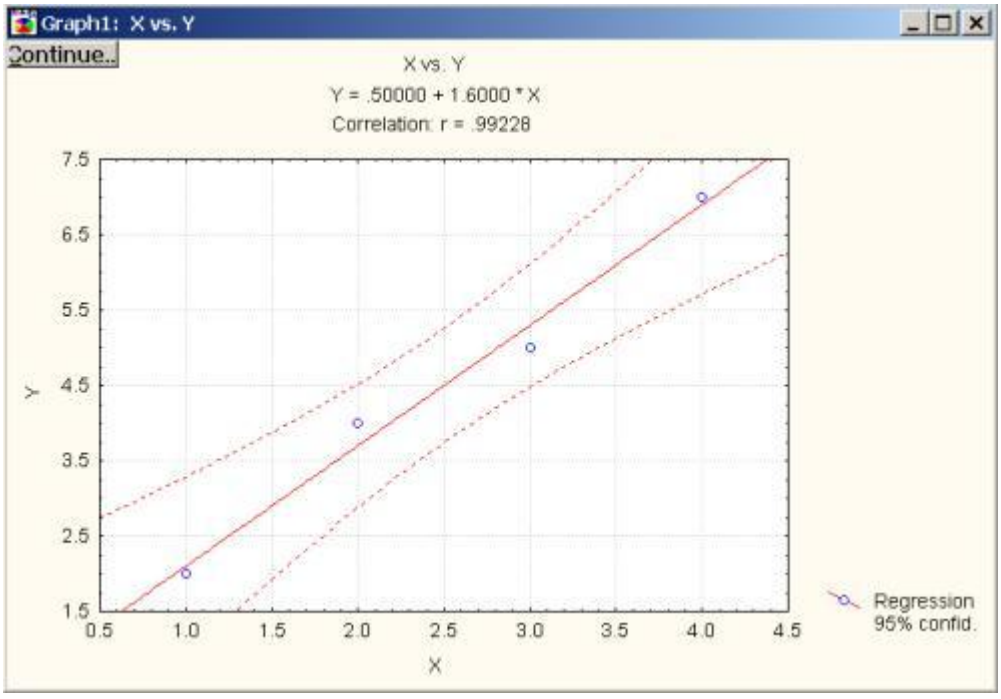


Рис. 9

Дальнейшие вычисления проводятся с использованием модуля **Multiple Regression**, для перехода в который щелкните дважды мышью в любом месте, свободном от окон. Появится стартовая панель Statistica Module Switcher. Выберите нужный модуль и нажмите кнопку **End & Switch To**. Появится окно **Multiple**

Regression. Нажмите кнопку **Variables** и определите зависимую (Dependent) и независимую (Independent) переменные (рис. 10).

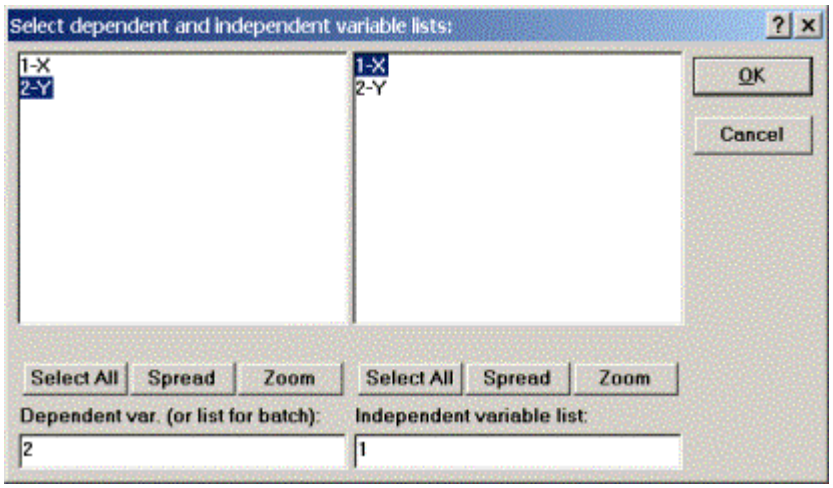


Рис. 10

- В окне **Multiple Regression** (рис.11) сделайте установки:

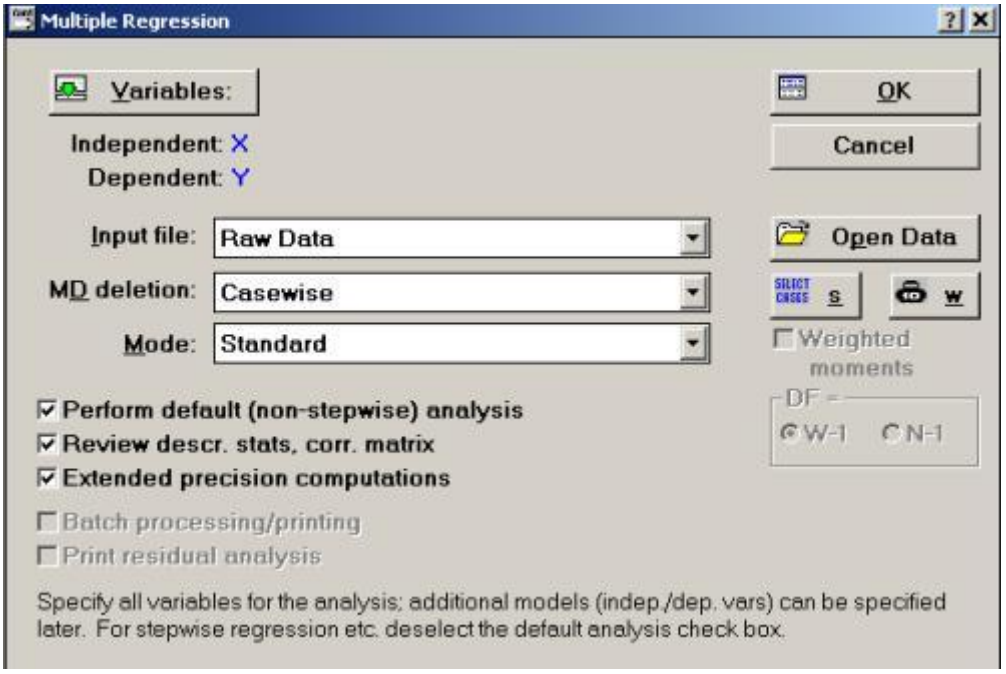


Рис. 11

- нажмите **OK**;
- появится окно **Review Descriptive Statistics** (рис. 12).

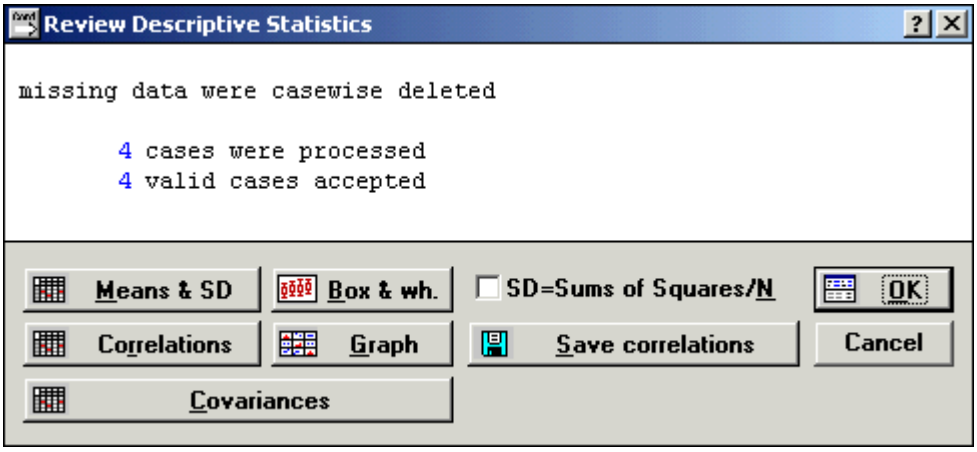
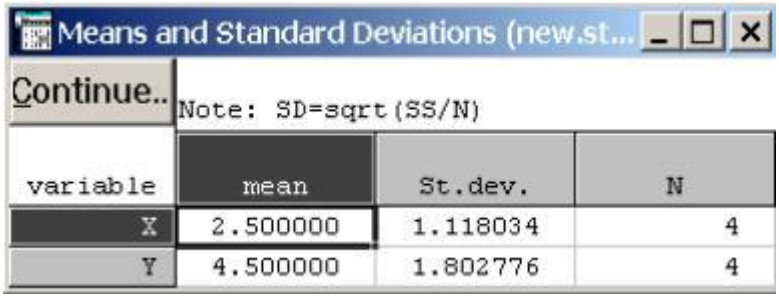


Рис. 12

• Определим смещенные среднеквадратические отклонения. Для этого поставьте флажок в поле **SD=Sums of Squares/N** и нажмите **Means & SD** (рис. 13). Выпишите необходимые данные. Аналогично можно вычислить несмещенные среднеквадратические отклонения при этом флажок должен отсутствовать (рис. 14).



variable	mean	St.dev.	N
X	2.500000	1.118034	4
Y	4.500000	1.802776	4

Рис. 13

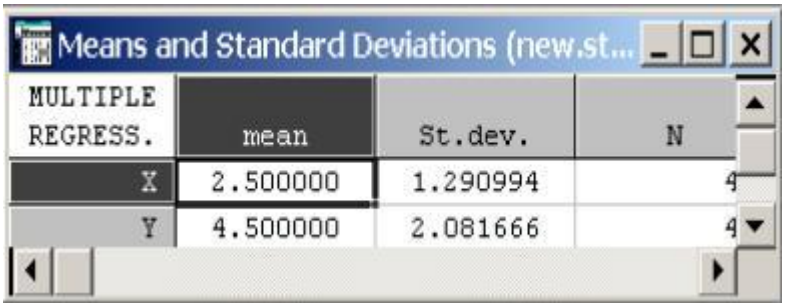


Рис. 14

• Нажмите кнопку **Continue**. Нажмите кнопку **Correlations** для расчета коэффициента корреляции (рис. 15).

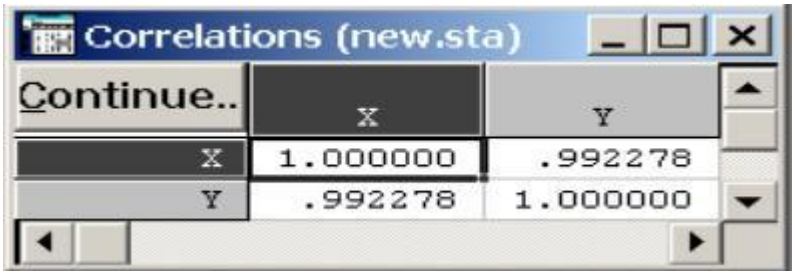


Рис. 15

- Выпишите необходимые данные.
- Нажмите кнопку **Continue**, затем **ОК**. Появится окно (рис. 16).

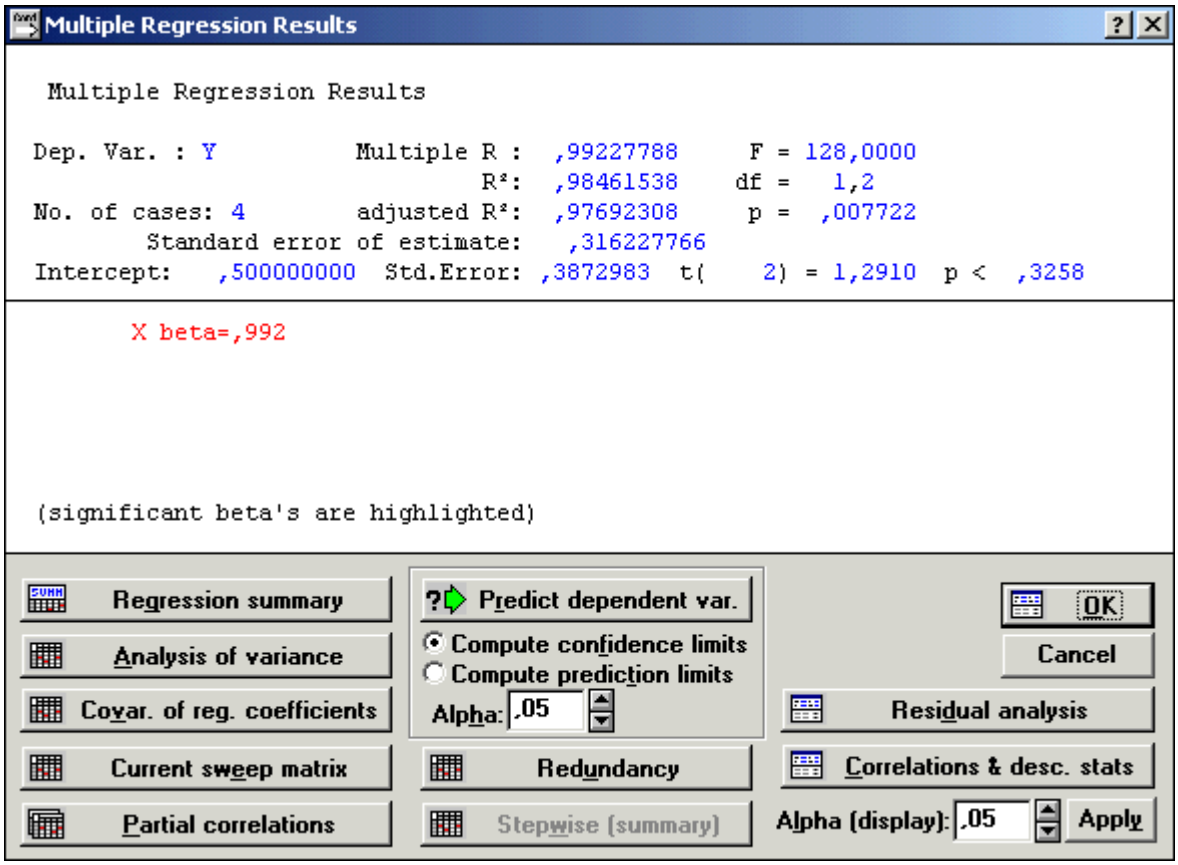
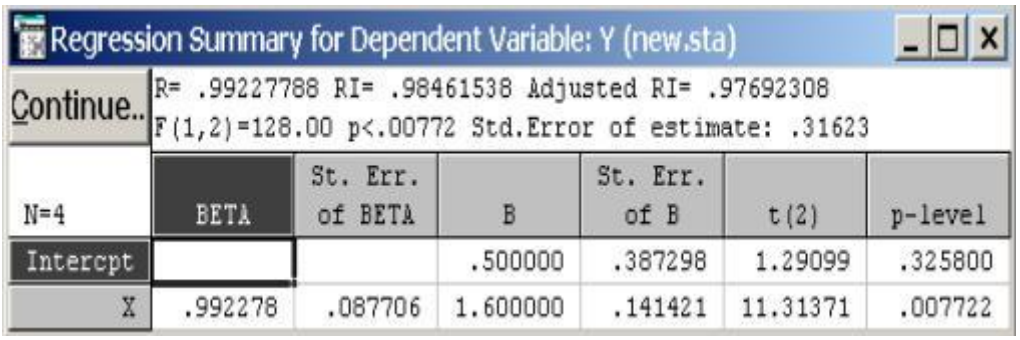


Рис. 16

- Нажмите кнопку Regression summary для расчета параметров модели и коэффициентов Стьюдента (рис. 17).



Regression Summary for Dependent Variable: Y (new.sta)

Continue..

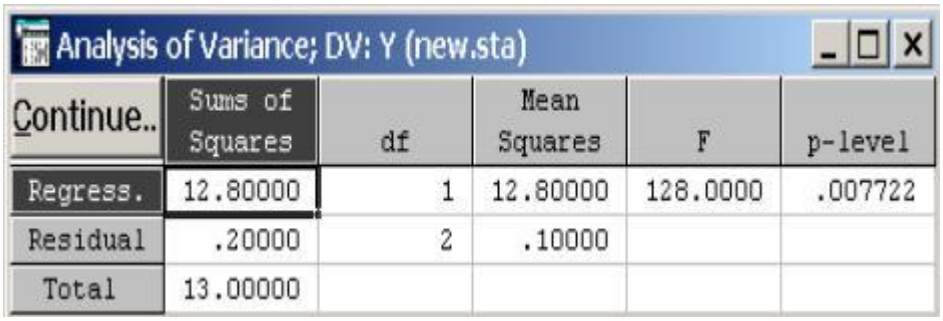
R= .99227788 RI= .98461538 Adjusted RI= .97692308
F(1,2)=128.00 p<.00772 Std.Error of estimate: .31623

N=4	BETA	St. Err. of BETA	B	St. Err. of B	t(2)	p-level
Intercpt			.500000	.387298	1.29099	.325800
X	.992278	.087706	1.600000	.141421	11.31371	.007722

Рис. 17

• Выпишите необходимые данные и нажмите кнопку **Continue**.

• Нажмите кнопку **Analysis of variance** для расчета таблицы адекватности и соответствующих QR , QE , Q, k, n–k–1, F, S (рис. 18).



Analysis of Variance; DV: Y (new.sta)

Continue..

	Sums of Squares	df	Mean Squares	F	p-level
Regress.	12.80000	1	12.80000	128.0000	.007722
Residual	.20000	2	.10000		
Total	13.00000				

Рис. 18

• Выпишите необходимые данные. Нажмите кнопку **Continue.** Нажмите кнопку **Residual analysis**. Появится окно (рис. 19).

• Нажмите кнопку **Display residuals & pred. (3)** для анализа остатков (рис. 20).

• Нажмите кнопку **Continue.** Нажмите кнопку **Durbin-Watson stat (4)** для расчета критерия Дарбина-Уотсона (рис. 21).

• Выпишите необходимые данные. Закройте окно программы STATISTICA.

• Проанализируйте полученные результаты, сделайте выводы о пригодности построенной математической модели для интерполяции и прогнозирования значений исследуемого показателя (критерий Фишера, критерий Стьюдента, коэффициент детерминации).

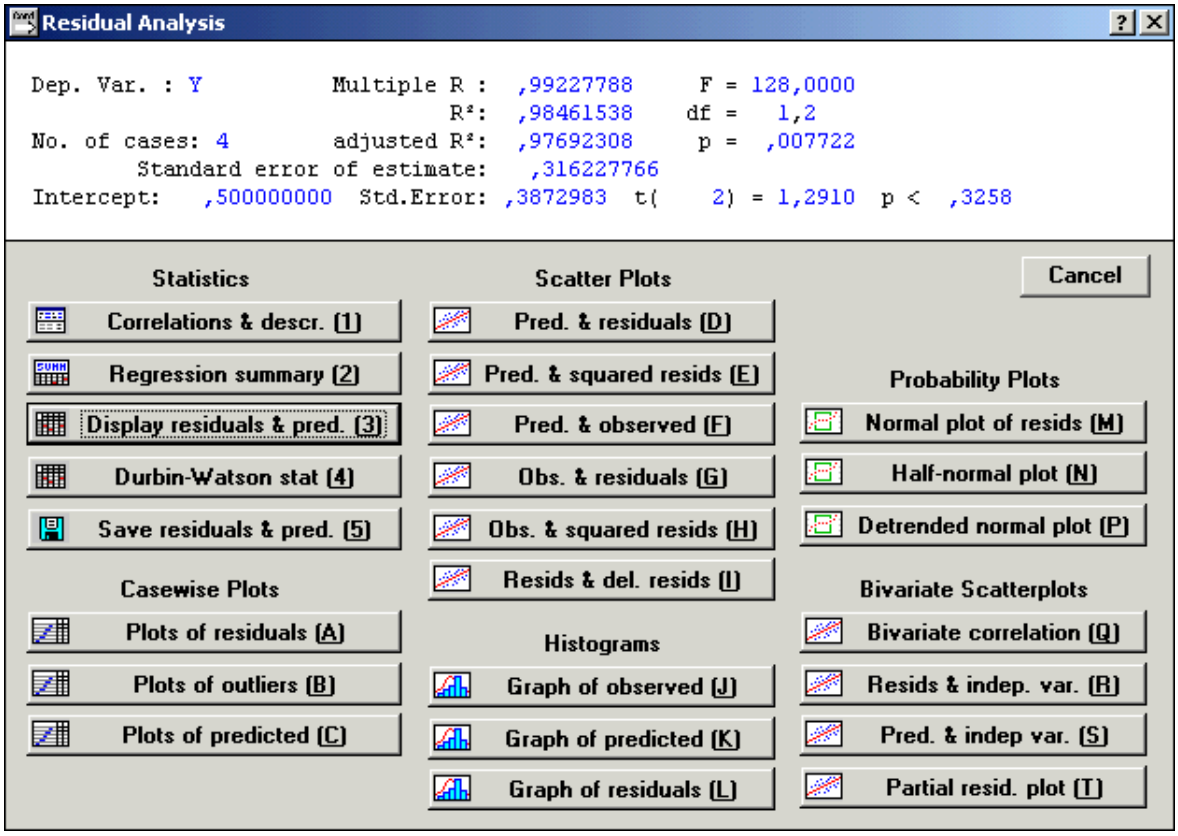
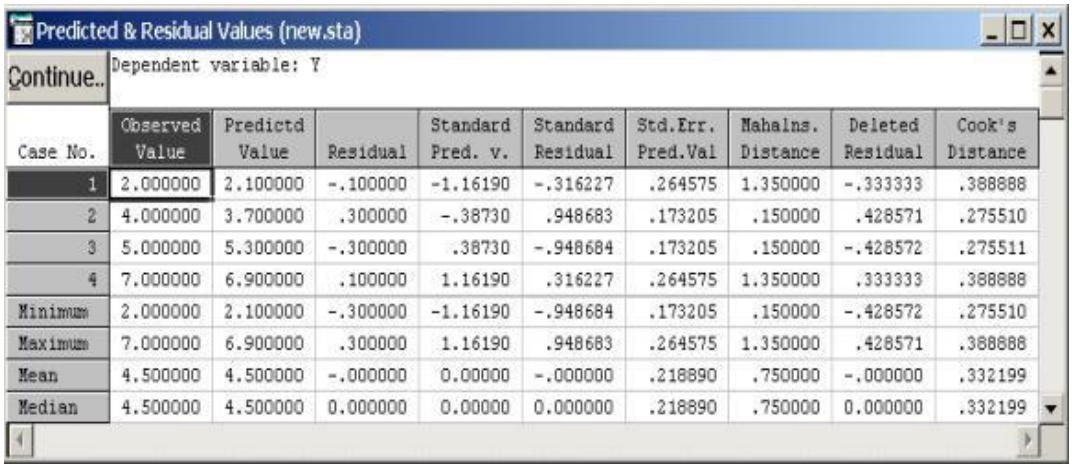
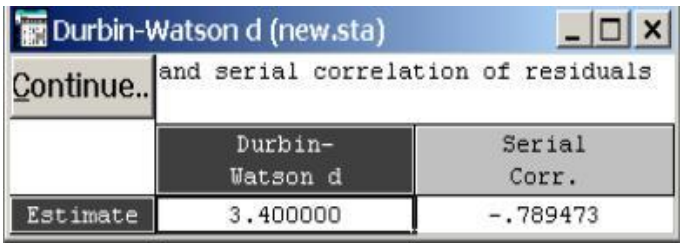


Рис. 19



Case No.	Observed Value	Predictd Value	Residual	Standard Pred. v.	Standard Residual	Std.Err. Pred.Val	Mahalns. Distance	Deleted Residual	Cook's Distance
1	2.000000	2.100000	-.100000	-1.16190	-.316227	.264575	1.350000	-.333333	.388888
2	4.000000	3.700000	.300000	-.38730	.948683	.173205	.150000	.428571	.275510
3	5.000000	5.300000	-.300000	.38730	-.948684	.173205	.150000	-.428572	.275511
4	7.000000	6.900000	.100000	1.16190	.316227	.264575	1.350000	.333333	.388888
Minimum	2.000000	2.100000	-.300000	-1.16190	-.948684	.173205	.150000	-.428572	.275510
Maximum	7.000000	6.900000	.300000	1.16190	.948683	.264575	1.350000	.428571	.388888
Mean	4.500000	4.500000	-.000000	0.00000	-.000000	.218890	.750000	-.000000	.332199
Median	4.500000	4.500000	0.000000	0.00000	0.000000	.218890	.750000	0.000000	.332199

Рис. 20



	Durbin-Watson d	Serial Corr.
Estimate	3.400000	-.789473

Рис. 21

4. Индивидуальные задания

- Для переменных, соответствующих вашему варианту (приложение 1), необходимо определить статистические характеристики (среднее значение, минимальное и максимальное значение, стандартное отклонение).
- Определите коэффициент парной корреляции и постройте соответствующий график.
- Постройте линейную регрессионную модель. Определите её адекватность (критерий Фишера), значимость параметров (критерий Стьюдента) и работоспособность (коэффициент детерминации).
- Постойте квадратичную регрессионную модель. Определите её адекватность, значимость параметров и работоспособность.
- Определите, какая регрессионная модель наиболее адекватная.

Таблица 5

Индивидуальные задания студентам для лабораторной работы по регрессионному анализу

Вариант 1	Вариант 7
X: 2.6 7.5 5.8 6.4 9.1	X: 0.1 3.8 5.6 7.0 6.9
Y: -1.8 0.4 0.1 -0.2 1.9	Y: -3.1 -0.7 -0.7 0.6 0.5
Вариант 2	Вариант 8
X: 1.2 0.1 6.2 8.1 5.4	X: 4.3 9.4 0.8 6.2 9.9
Y: -2.8 -3.5 0.1 0.4 0.0	Y: -0.9 2.2 -3.0 0.7 2.2
Вариант 3	Вариант 9
X: 0.0 3.9 6.5 3.2 1.9	X : 8.7 6.4 6.0 8.7 3.1
Y: -3.0 -0.6 0.1 -1.0 -2.1	Y: 1.3 0.4 0.5 1.1 -2.2
Вариант 4	Вариант 10
6.6 8.9 2.7 3.1 8.8	3.1 3.3 1.2 1.2 6.4
0.2 1.5 -1.5 -1.1 1.2	-1.1 -0.6 -2.6 -3.1 0.0
Вариант 5	Вариант 11
5.4 2.7 3.1 8.1 5.3	7.3 8.9 6.4 0.4 9.4
0.0 -1.3 -1.1 1.4 -0.6	0.7 1.0 0.4 -2.1 2.1
Вариант 6	Вариант 12
4.0 5.2 3.5 9.8 1.6	1.7 5.8 1.6 2.9 2.6
-0.2 -0.7 -1.6 1.9 -2.1	-2.1 0.8 -2.3 -2.1 -1.5

Продолжение табл. 5

Вариант 13	Вариант 20
2.6 7.8 3.9 1.5 7.8	6.1 2.7 6.7 5.4 5.9
-2.5 1.2 -0.8 -2.1 1.4	0.9 -2.2 0.0 -0.9 -0.4
Вариант 14	Вариант 21
1.4 1.6 4.2 6.6 4.3	0.6 9.6 1.9 8.0 9.1
-2.3 -1.7 -0.8 -0.1 -0.7	-2.3 2.2 -2.7 0.3 1.8
Вариант 15	Вариант 22
0.0 4.2 4.6 8.3 0.5	2.7 7.7 3.0 5.3 2.8
-3.3 -0.6 -0.8 1.7 -2.6	-2.2 0.3 -1.2 0.2 -1.3
Вариант 16	Вариант 23
8.9 8.0 5.0 3.3 7.0	7.1 4.6 8.1 7.9 6.1
1.9 0.7 0.0 -1.3 0.3	0.0 -1.1 1.5 1.1 -0.3
Вариант 17	Вариант 24
7.5 0.5 5.4 5.0 3.3	1.5 1.5 3.3 0.4 9.3
0.5 -2.4 0.4 -0.3 -1.6	-2.4 -1.8 -2.0 -3.5 2.3
Вариант 18	Вариант 25
6.1 3.1 5.8 6.7 9.5	5.7 7.1 8.6 9.6 2.3
-0.1 -1.8 0.8 0.2 1.9	-0.2 1.5 1.5 2.0 -1.4
Вариант 19	Вариант 26
2.8 9.3 1.9 10.0 6.7	0.1 4.1 3.7 2.1 5.6
-0.7 1.8 -2.0 1.6 -0.9	-3.0 -1.1 -0.8 -2.2 -0.4

ПРИЛОЖЕНИЕ 1

Распределение Фишера.
Значения квантилей для степеней свободы f1 и f2 и вероятности λ=0,05

$f_2 \backslash f_1$	1	2	3	4	5	6	8	12	24
1	161.45	199.50	215.72	224.57	230.17	233.97	238.89	243.91	249.04
2	18.512	18.999	19.163	19.248	19.298	19.329	19.371	19.414	19.453
3	10.129	9.552	9.276	9.118	9.014	8.941	8.844	8.744	8.638
4	7.710	6.945	6.591	6.388	6.257	6.164	6.041	5.912	5.774
5	6.607	5.786	5.410	5.192	5.050	4.950	4.818	4.678	4.527
6	5.987	5.143	4.756	4.388	4.284	4.147	4.000	3.841	3.669
7	5.591	4.737	4.347	4.121	3.972	3.866	3.725	3.574	3.410
8	5.317	4.459	4.067	3.838	3.688	3.580	3.438	3.284	3.116
9	5.117	4.256	3.863	3.633	3.482	3.374	3.230	3.073	2.900
10	4.965	4.103	3.708	3.478	3.326	3.217	3.072	2.913	2.737
11	4.844	3.982	3.587	3.357	3.204	3.094	2.948	2.778	2.609
12	4.747	3.885	3.490	3.259	3.106	2.999	2.848	2.686	2.505
13	4.667	3.805	3.410	3.179	3.025	2.915	2.767	2.604	2.420
14	4.600	3.739	3.344	3.112	2.958	2.848	2.699	2.534	2.349
15	4.543	3.683	3.287	3.056	2.901	2.790	2.641	2.475	2.288
16	4.494	3.634	3.239	3.007	2.853	2.741	2.591	2.424	2.235
17	4.451	3.592	3.197	2.965	2.810	2.699	2.548	2.381	2.190
18	4.414	3.555	3.160	2.928	2.773	2.661	2.510	2.342	2.150
19	4.381	3.522	3.127	2.895	2.740	2.629	2.477	2.308	2.114
20	4.351	3.493	3.098	2.866	2.711	2.599	2.447	2.278	2.083
21	4.325	3.467	3.072	2.840	2.685	2.573	2.421	2.250	2.054
22	4.301	3.443	3.049	2.817	2.661	2.549	2.397	2.226	2.028
23	4.279	3.422	3.028	2.795	2.640	2.528	2.375	2.203	2.005
24	4.260	3.403	3.009	2.777	2.621	2.508	2.355	2.183	1.984
25	4.242	3.385	2.991	2.759	2.603	2.490	2.337	2.165	1.965
26	4.225	3.369	2.975	2.743	2.587	2.474	2.321	2.148	1.947
27	4.210	3.354	2.961	2.728	2.572	2.459	2.305	2.132	1.930
28	4.196	3.340	2.947	2.714	2.558	2.445	2.292	2.118	1.915
29	4.183	3.328	2.934	2.702	2.545	2.432	2.278	2.104	1.901
30	4.171	3.316	2.922	2.690	2.534	2.421	2.266	2.092	1.887
60	4.001	3.151	2.758	2.525	2.368	2.254	2.097	1,918	1.700
120	3.920	3.072	2.680	2.447	2.290	2.175	2.016	1.834	1.608

ПРИЛОЖЕНИЕ 2

Распределение Стьюдента.
Значения квантилей для степеней свободы f и вероятности λ=0,05

f	1	2	3	4	5	6	7	8	9
t	12.71	4.303	3.182	2.776	2.571	2.447	2.365	2.306	2.262
f	10	11	12	13	14	15	16	17	18
t	2.228	2.201	2.179	2.160	2.145	2.131	2.120	2.110	2.101
f	19	20	21	22	23	24	25	26	27
t	2.093	2.086	2.080	2.074	2.069	2.064	2.060	2.056	2.052
f	28	29	30	40	50	60	80	100	200
t	2.048	2.045	2.042	2.021	2.009	2.000	1.990	1.984	1.972

ПРИЛОЖЕНИЕ 3

Критерий Дарбина-Уотсона.
Нижние и верхние границы критерия для вероятности λ=0,05

F_1	f_2=1	f_2=2	f_2=3	f_2=4	f_2=5	f_2=6	f_2=7
	dн dв	dн dв	dн dв	dн dв	dн dв	dн dв	dн dв
6	0.61 1.40						
7	0.70 1.36	0.47 1.90					
8	0.76 1.33	0.56 1.78	0.37 2.29				
9	0.82 1.32	0.63 1.70	0.46 2.13	0.30 2.59			
10	0.88 1.32	0.70 1.64	1.53 2.02	0.38 2.41	0.24 2.82		
11	0.93 1.32	0.76 1.60	0.60 1.93	0.44 2.28	0.32 2.65	0.20 3.01	
12	0.97 1.33	0.81 1.58	0.66 1.86	0.51 2.18	0.38 2.51	0.27 2.83	0.17 3.15
13	1.01 1.34	0.86 1.56	0.72 1.82	0.57 2.09	0.45 2.39	0.33 2.69	0.23 2.99
14	1.05 1.35	0.91 1.55	0.77 1.78	0.63 2.03	0.51 2.30	0.39 2.57	0.29 2.85
15	1.08 1.36	0.95 1.54	0.81 1.75	0.69 1.98	0.56 2.22	0.45 2.47	0.34 2.73
16	1.11 1.37	0.98 1.54	0.86 1.73	0.73 1.94	0.62 2.16	0.50 2.39	0.40 2.62
17	1.13 1.38	1.02 1.54	0.90 1.71	0.78 1.90	0.66 2.10	0.55 2.32	0.45 2.54
18	1.16 1.39	1.05 1.54	0.93 1.70	0.82 1.87	0.71 2.06	0.60 2.26	0.50 2.46
19	1.18 1.40	1.07 1.54	0.97 1.69	0.86 1.85	0.75 2.02	0.65 2.21	0.55 2.40
20	1.20 1.41	1.10 1.54	1.00 1.68	0.89 1.83	0.79 1.99	0.69 2.16	0.60 2.34
25	1.29 1.45	1.21 1.55	1.12 1.65	1.04 1.77	0.95 1.89	0.87 2.01	0.78 2.14
30	1.35 1.49	1.28 1.57	1.21 1.65	1.14 1.74	1.07 1.83	1.00 1.93	0.93 2.03
40	1.44 1.54	1.39 1.60	1.34 1.66	1.29 1.72	1.23 1.79	1.18 1.85	1.12 1.92
50	1.50 1.59	1.46 1.63	1.42 1.67	1.38 1.72	1.34 1.77	1.29 1.82	1.25 1.88
60	1.55 1.62	1.51 1.65	1.48 1.69	1.44 1.73	1.41 1.77	1.37 1.81	1.34 1.85
70	1.58 1.64	1.55 1.67	1.53 1.70	1.49 1.74	1.46 1.77	1.43 1.80	1.40 1.84
80	1.61 1.66	1.59 1.69	1.56 1.72	1.53 1.74	1.51 1.77	1.48 1.80	1.45 1.83
90	1.64 1.68	1.61 1.69	1.59 1.73	1.57 1.75	1.54 1.78	1.52 1.80	1.49 1.83
100	1.65 1.69	1.63 1.70	1.61 1.74	1.59 1.76	1.57 1.78	1.55 1.80	1.53 1.83
150	1.72 1.75	1.71 1.76	1.69 1.77	1.68 1.79	1.67 1.80	1.65 1.82	1.64 1.83
200	1.76 1.78	1.75 1.79	1.74 1.80	1.73 1.81	1.72 1.82	1.71 1.83	1.70 1.84

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Айвазян С.А. Прикладная статистика. Основы моделирования и первичная обработка данных / С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин. М.: Финансы и статистика,1983.

2. Коровин Е.Н. Методы обработки биомедицинских данных: учеб. пособие / Е.Н. Коровин, О.В. Родионов. Воронеж: ГОУВПО «Воронежский государственный технический университет», 2007.

3. Фестер Э. Методы корреляционного и регрессионного анализа / Э. Фестер, В. Ренц. М.: Финансы и статистика, 1983.