

КАЛИНИНГРАДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

**СТАТИСТИКА
(КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ
АНАЛИЗ СТАТИСТИЧЕСКИХ СВЯЗЕЙ
НА ПЕРСОНАЛЬНОМ КОМПЬЮТЕРЕ)**

**Методические указания
к практическим занятиям
для студентов всех форм обучения
специальности «Менеджмент»**

**Калининград
1999**

Статистика: Корреляционно-регрессионный анализ статистических связей на персональном компьютере: Методические указания к практическим занятиям для студентов всех форм обучения специальности «Менеджмент» / Калинингр. ун-т; Сост. Н.Ю. Лукьянова. - Калининград, 1999. - 35 с.

Методические указания разработаны в соответствии с учебным планом специальности «Менеджмент»; содержат основные теоретические положения корреляционно-регрессионного анализа, общие рекомендации по автоматизированному решению соответствующих задач, вопросы для самопроверки, список рекомендуемой литературы.

Составитель: канд. экон. наук, ст. преподаватель Н.Ю. Лукьянова.

СТАТИСТИКА

Корреляционно-регрессионный анализ
статистических связей на персональном компьютере

Методические указания
к практическим занятиям для студентов всех
форм обучения специальности «Менеджмент»

Составитель Наталия Юрьевна Лукьянова

Лицензия № 020345 от 14.01.1997 г.

Редактор Л.Г. Ванцева.

Подписано в печать 23.04.1999 г. Формат 60×90 ¹/₁₆.

Гарнитура «Таймс». Бумага для множительных аппаратов. Ризограф.

Усл. печ. л. 2,1. Уч.-изд. л. 2,0. Тираж 200 экз. Заказ .

Калининградский государственный университет
236041, г. Калининград, ул. А. Невского, 14

СОДЕРЖАНИЕ

Введение	4
1. Краткий обзор статистических программных продуктов	5
2. Основные теоретические положения корреляционно-регрессионного анализа статистических связей	8
2.1. Парная корреляция и регрессия	9
2.2. Множественная корреляция и регрессия	12
3. Решение задач корреляционно-регрессионного анализа статистических связей признаков на персональном компьютере в среде пакета <i>STATISTICA</i>	16
3.1. Общие сведения об интегрированном статистическом пакете общего назначения <i>STATISTICA</i>	16
3.2. Пример решения задачи	21
3.3. Порядок выполнения индивидуального задания	28
4. Вопросы для самопроверки	29
Список рекомендуемой литературы	30
Приложение 1. Таблица значений F-критерия Фишера	32
Приложение 2. Значения t-критерия Стьюдента	34

ВВЕДЕНИЕ

В условиях рыночной конкуренции процесс подготовки и принятия решений менеджерами компаний должен включать тщательный анализ имеющихся данных, базирующийся на методах математической статистики. В этой связи существенную помощь в получении необходимой информации могут оказать современные информационные технологии интеллектуального и статистического анализа данных. Оценка кредитных и страховых рисков, прогнозирование тенденций на финансовых рынках, оценка объектов недвижимости, построение профилей потенциальных покупателей определенного товара, анализ продуктовой корзины - вот далеко не полный перечень задач, успешно решаемых с помощью систем интеллектуального и статистического анализа данных.

Системы интеллектуального анализа предназначены для автоматизированного поиска ранее неизвестных закономерностей в имеющихся в распоряжении менеджера данных с последующим использованием полученной информации для подготовки решений. Помимо статистических методов базовыми инструментами анализа в таких системах являются нейронные сети, деревья решений и индукция правил. Однако несмотря на то, что в последние годы рынок программных продуктов этого типа активно развивается, они все еще недоступны по цене предприятиям среднего и малого бизнеса. В то же время компаниям такого размера, как правило, не требуется столь мощный аналитический инструментарий, предлагаемый этими системами.

Более доступными средствами анализа данных на сегодняшний день являются статистические программные продукты (СПП). В мировой практике компьютерные системы статистического анализа и обработки данных широко применяются как в исследовательской работе в области экономики, так и в практической деятельности аналитических, маркетинговых и плановых отделов банков, страховых компаний, производственных и торговых фирм. В последние годы заметно возрос спрос на СПП и в нашей стране.

СПП позволяют решить широкий спектр задач «разведочного» анализа данных, статистического исследования зависимостей, планирования экспериментов, анализа временных рядов, анализа данных нечисловой природы и т.д. Настоящие методические разработки посвящены вопросам корреляционно-регрессионного анализа статистических связей с использованием одного из самых популярных в России статистических программных продуктов - пакета *STATISTICA*, функционирующего в среде Windows.

1. КРАТКИЙ ОБЗОР СТАТИСТИЧЕСКИХ ПРОГРАММНЫХ ПРОДУКТОВ

Рынок СПП необычайно разнообразен. Существует около тысячи распространяемых на мировом рынке пакетов, решающих задачи статистического анализа данных в среде DOS, OS/2 или Windows. Можно выделить четыре основные группы статистических пакетов (рис.1).



Рис. 1. Основные группы статистических программных продуктов

Остановимся подробнее на методоориентированных пакетах (табл. 1) [1, 2].

Таблица 1

Классификация методоориентированных статистических программ

Класс статистических программных продуктов	Наименование статистических программных продуктов
Универсальные (интегрированные) статистические пакеты общего назначения Инструментарий для исследователей, включающий мощную статистическую компоненту	SAS, SPSS для Windows, SYSTAT, MINITAB, Statgraphics, BMDP Dynamic, STATISTICA/W, Stat View и Super ANONA IMSL, S-Plus
Специализированные пакеты по классификации и снижению размерности	КЛАСС-МАСТЕР, Stat-Media, PALMODA (ЛОРЕГ), STARC, КВАЗАР, PolyAnalyst, MVSP, CART

Класс статистических программных продуктов	Наименование статистических программных продуктов
Некоторые другие специализированные и универсальные СПП	МЕЗОЗАВР (MESOSAUR), САНИ (SANI), Stat View for Windows, STADIA, ОЛИМП, РОСТАН, NCSS Statistical Software, ODA, SOLO, STATlab Pro, UNISTAT, STATIT, WinSTAT, Multivariate 7, JMP, BM-STAT, DATA DESK, SAM-86, STATMOST, POWERSTAT
Пакеты и программы, решающие смежные с классификацией задачи	«Статистик-Консультант», BMDP для Windows, TURBO Spring-Stat-Win, STATISTIX, SigmaStat, StatXact-3, MS-Excel-5.0
Статистические экспертные системы	СТАТЭКС, Statistical Navigator Pro, STAREX

В **универсальных пакетах**, предлагающих широкий диапазон статистических методов, отсутствует ориентация на конкретную предметную область. Из зарубежных универсальных пакетов наибольшую известность получили компьютерные системы SAS, SPSS, SYSTAT, Minitab, Statgraphics, *Statistica*.

Специализированные пакеты, как правило, содержат несколько статистических методов или методы, применяемые в конкретной предметной области. Чаще всего это системы, ориентированные на анализ временных рядов, корреляционно-регрессионный, факторный или кластерный анализ. **«Полуспециализированными» и «полууниверсальными»** можно считать российские пакеты STADIA, ОЛИМП и белорусский пакет РОСТАН. К этому же классу следует отнести и американские пакеты ODA, WinSTAT, Statit, UNISTAT, Multivariate 7, JMP, SOLO, STATlab. К **специализированным пакетам по классификации и снижению размерности** можно отнести такие отечественные системы, как КЛАСС-МАСТЕР, КВАЗАР, PALMODA, Stat-Media, STARC, а также ряд зарубежных пакетов, например MVSP.

Широко известны пакеты, решающие смежные с классификацией задачи: американские системы BMDP/W, SigmaStat, Statistix, TURBO Spring-Stat-Win, а также отечественный пакет «Статистик-Консультант для Windows». Кроме того, на рынке имеются **статистические экспертные системы**, например СТАТЭКС, Statistical Navigator Pro. Среди нестатистических пакетов, решающих задачи классификации, можно отметить пакеты PolyAnalyst, ДА-система, АРГОНАВТ, ЛОРЕГ, пакет ОТЭКС и разнообразные нейросетевые пакеты.

В состав методоориентированных СПП могут входить следующие функциональные блоки.

I. Блок описательной статистики и разведочного анализа исходных данных предусматривает:

- анализ смешанной природы многомерного признака и унификацию записи исходных данных;
- анализ резко выделяющихся наблюдений;
- восстановление пропущенных наблюдений;
- проверку статистической независимости наблюдений;
- определение основных числовых характеристик и частотную обработку исходных данных (построение гистограмм, полигонов частот, вычисление выборочных средних, дисперсий);
- статистическое оценивание параметров;
- вычисление модельных законов распределения вероятностей (нормального, биномиального, Пуассона, хи-квадрат и др.);
- визуализацию анализируемых многомерных статистических данных и др.

II. Блок статистического исследования зависимостей предполагает:

- корреляционно-регрессионный анализ;
- дисперсионный и ковариационный анализ;
- планирование регрессионных экспериментов и выборочных обследований;
- анализ временных рядов (предварительный анализ временных рядов; выявление тренда временного ряда; выявление скрытых периодичностей, спектральный анализ временного ряда, анализ случайных остатков временного ряда; проверка статистических гипотез: о стационарности ряда, о независимости его членов, об адекватности «подгоняемой» модели) и др.

III. Блок классификации и снижения размерности включает:

- дискриминантный анализ;
- статистический анализ смесей распределений;
- кластер-анализ;
- снижение размерности в соответствии с критериями внешней информативности и автоинформативности и некоторые др.

IV. Блок методов статистического анализа нечисловой информации и экспертных оценок. Среди используемого в этом блоке математико-статистического инструментария - анализ таблиц сопряженности, логлинейные модели, субъективные вероятности, логит- и пробит-анализ, ранговые методы и т.п.

V. Блок планирования эксперимента и выборочных обследований.

VI. Блок вспомогательных программ предусматривает статистическое моделирование на ЭВМ, включая генерирование одномерных и

многомерных наблюдений, «извлеченных» из генеральных совокупностей заданного типа.

Одним из наиболее динамично развивающихся универсальных методоориентированных статистических пакетов является система Statistica для Windows (далее *STATISTICA*) американской фирмы StatSoft (<http://www.statsoft.com>). По результатам многочисленных рейтингов *STATISTICA* стала мировым лидером на рынке СПП и вошла в число 100 лучших программных продуктов (Windows Magazin, февраль 1995), а также занимает первое место среди СПП по результатам последнего рейтинга (BYTE, сентябрь 1998).

2. ОСНОВНЫЕ ТЕОРЕТИЧЕСКИЕ ПОЛОЖЕНИЯ КОРРЕЛЯЦИОННО-РЕГРЕССИОННОГО АНАЛИЗА СТАТИСТИЧЕСКИХ СВЯЗЕЙ

Существует два основных типа связей между социально-экономическими явлениями и их признаками: **функциональная** (жестко детерминированная) и **статистическая** (стохастически детерминированная). При функциональной связи каждому значению факторного признака соответствуют строго определенные значения результативного признака. При статистической связи с изменением значения факторного признака значения результативного признака могут варьировать в определенных пределах, т.е. принимать любые значения в этих пределах с некоторыми вероятностями. При этом статистические характеристики результативного признака изменяются по определенному закону. Статистическая связь проявляется не в каждом отдельном случае, а в среднем. **Корреляционная связь** (англ. correlation - соответствие) является частным случаем статистической связи, при которой изменение среднего значения результативного признака обусловлено изменением значений факторного признака (*парная корреляция*) или множества факторных признаков (*множественная корреляция*). Для оценки тесноты связи (связь отсутствует, слабая, умеренная, сильная), определения ее направленности (связь прямая или обратная), а также формы (связь линейная, параболическая, гиперболическая, степенная и т.д.) используется корреляционно-регрессионный метод.

Корреляционно-регрессионный анализ позволяет количественно измерить тесноту, направление связи (корреляционный анализ), а также установить аналитическое выражение зависимости результата от конкретных факторов при постоянстве остальных действующих на результативный признак факторных признаков (регрессионный анализ).

Основные условия применения корреляционно-регрессионного метода

1. Наличие достаточно большой по объему выборочной совокупности. Считается, что число наблюдений должно превышать более чем в 10 раз число факторов, влияющих на результат.
2. Наличие качественно однородной исследуемой совокупности.
3. Подчинение распределения совокупности по результативному и факторным признакам нормальному закону или близость к нему. Выполнение этого условия обусловлено использованием **метода наименьших квадратов** (МНК) при расчете параметров корреляции (см. п. 2.1) и некоторых др.

Основные задачи корреляционно-регрессионного анализа

1. *Измерение тесноты связи между результативным и факторным признаком (признаками).* В зависимости от количества влияющих на результат факторов задача решается путем вычисления корреляционного отношения, коэффициентов парной, частной, множественной корреляции или детерминации.
2. *Оценка параметров уравнения регрессии,* выражающего зависимость средних значений результативного признака от значений факторного признака (признаков). Задача решается путем вычисления коэффициентов регрессии.
3. *Определение важнейших факторов, влияющих на результативный признак.* Задача решается путем оценки тесноты связи факторов с результатом.
4. *Прогнозирование возможных значений результативного признака при задаваемых значениях факторных признаков.* Задача решается путем подстановки ожидаемых значений факторов в регрессионное уравнение и вычисления прогнозируемых значений результата.

2.1. Парная корреляция и регрессия

Часто при анализе взаимосвязей социально-экономических явлений среди различных факторов, влияющих на результат, бывает важно выделить наиболее значимый факторный признак, который в большей степени обуславливает вариацию результативного признака (например, зависимость проданных туристическими фирмами путевок от затрат на рекламу или зависимость производительности труда операторов ЭВМ от стажа работы). Этим обусловлена необходимость измерения парных корреляций и построения уравнений парных регрессий.

Парная корреляция характеризует тесноту и направленность связи между результативным и факторным признаками. *Парная регрессия* позволяет описать форму связи в виде уравнения парной регрессии (табл.2).

Таблица 2

Основные виды уравнений парной регрессии

Наименование формы парной регрессии	Вид уравнения парной регрессии
Линейная	$\tilde{y} = a_0 + a_1 x$
Гиперболическая	$\tilde{y} = a_0 + a_1 (1/x)$
Параболическая	$\tilde{y} = a_0 + a_1 x + a_2 x^2$
Степенная	$\tilde{y} = a_0 x^{a_1}$

В данной таблице \tilde{y} – теоретическое значение результативного признака (y) при определенном значении факторного признака (x), подставленном в регрессионное уравнение; a_0 – свободный член уравнения; a_1, a_2 – коэффициенты регрессии.

Параметры уравнений парной регрессии a_1, a_2 называют *коэффициентами регрессии*. Для оценки параметров уравнения парной регрессии используется метод наименьших квадратов (МНК). Он заключается в определении параметров a_0, a_1, a_2 , при которых сумма квадратов отклонений фактических значений результата (y_i) от теоретических (\tilde{y}_i) минимизируется. Так, (2.1) описывает исходное условие МНК для парной линейной корреляционной связи:

$$\sum_{i=1}^n (y_i - \tilde{y}_i)^2 \rightarrow \min, \quad \text{или} \quad (2.1)$$

$$f(a_0, a_1) = \sum_{i=1}^n [y_i - (a_0 + a_1 x_i)]^2 \rightarrow \min.$$

На основе (2.1) определяются частные производные функции $f(a_0, a_1)$, которые затем приравниваются к 0. Далее полученные уравнения преобразуются в систему нормальных уравнений, из которых определяются параметры a_0, a_1 . При этом число нормальных уравнений в общем случае будет равно числу параметров. При использовании СПП параметры регрессионного уравнения определяются автоматически. Подробнее МНК изложен в [6, 7].

В частности, коэффициент парной линейной регрессии a_1 определяется в соответствии с (2.2.) и характеризует меру связи между вариациями факторного и результативного признаков. Коэффициент регрессии показывает, на сколько в среднем изменяется значение результативного признака при изменении факторного на единицу:

$$a_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (2.2)$$

где n – объем совокупности.

Тесноту и направление парной линейной корреляционной связи измеряют с помощью *линейного коэффициента корреляции* (2.3), принимающего значения в пределах от -1 до $+1$ (см. табл.3):

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}. \quad (2.3)$$

Квадрат коэффициента корреляции называют *коэффициентом детерминации* (r^2). Коэффициент детерминации можно интерпретировать как долю общей дисперсии результативного признака (y), которая объясняется вариацией факторного признака (x).

Таблица 3

Оценка характера связи по линейному коэффициенту корреляции

Значения линейного коэффициента корреляции	Характер связи
$r = -1$	функциональная
$-1 < r < -0,7$	обратная сильная
$-0,7 \leq r \leq -0,5$	обратная умеренная
$-0,5 < r < 0$	обратная слабая
$r = 0$	отсутствует
$0 < r < +0,5$	прямая слабая
$+0,5 \leq r \leq +0,7$	прямая умеренная
$+0,7 < r < +1$	прямая сильная
$r = +1$	функциональная

Значимость линейного коэффициента корреляции проверяется на основе t-критерия Стьюдента: проверяется нулевая гипотеза об отсутствии связи между факторным и результативным признаками ($H_0: r = 0$). Для проверки H_0 по формуле (2.4) следует рассчитать t-статистику (t_p) и сравнить ее с табличным значением (t_r), определяемым с использованием

таблицы приложения 2 по заданным уровню значимости (α) и числу степеней свободы ($d.f.$). Если $t_p > t_r$, то гипотеза H_0 отвергается с вероятностью ошибки меньше чем $\alpha \cdot 100\%$. Это свидетельствует о значимости линейного коэффициента корреляции и статистической существенности зависимости между факторным и результативным признаками.

$$t_p = \frac{|r| \sqrt{k}}{\sqrt{1-r^2}}, \quad (2.4)$$

где $k = n-2$ для малой выборки,

$k = n$ при большом числе наблюдений ($n > 100$).

Аналогично оценивается значимость коэффициента регрессии; t_p рассчитывают как отношение взятого по модулю коэффициента регрессии к его средней ошибке с заданными уровнем значимости (α) и числом степеней свободы $d.f. = n-2$.

2.2. Множественная корреляция и регрессия

При анализе взаимосвязей социально-экономических явлений, как правило, выясняется, что на результат влияет ряд факторных признаков, основные из которых следует включить в регрессионную модель. При этом следует помнить, что все факторы учесть в модели невозможно по ряду причин: часть факторов просто неизвестна современной науке, по части известных факторов нет достоверной информации или количество включаемых в модель факторов может быть ограничено объемом выборки (количество факторных признаков должно быть на порядок меньше численности изучаемой совокупности).

Множественная регрессия описывает форму связи в виде уравнения множественной регрессии, или регрессионной модели (табл.4).

Таблица 4

Основные виды множественной регрессии

Форма регрессии	Вид уравнения регрессии
Линейная	$\tilde{y} = a_0 + a_1 x_1 + \dots + a_m x_m$
Гиперболическая	$\tilde{y} = a_0 + a_1 (1/x_1) + \dots + a_m (1/x_m)$
Параболическая	$\tilde{y} = a_0 + a_1 x_1^2 + \dots + a_m x_m^2$
Степенная	$\tilde{y} = a_0 x_1^{a_1} x_2^{a_2} \dots x_m^{a_m}$

\tilde{y} – теоретическое значение результативного признака (y) при определенных значениях факторных признаков (x_1, x_2, \dots, x_m), подставленных в регрессионное уравнение;

a_0 – свободный член уравнения;

a_1, a_2, \dots, a_m – коэффициенты множественной регрессии.

Параметры уравнения множественной регрессии a_1, a_2, \dots, a_m называют *коэффициентами множественной регрессии* и определяют с помощью МНК путем решения системы нормальных уравнений МНК. При этом число нормальных уравнений в общем случае будет равно числу параметров. Если связь отдельного фактора с результатом не является линейной, то производят линеаризацию уравнения. Для упрощения решения системы нормальных уравнений значения всех признаков заменяют на отклонения индивидуальных значений признаков от их средних величин. Полученные коэффициенты множественной регрессии являются именованными числами и показывают, на сколько изменится результативный признак (по отношению к своей средней величине) при отклонении факторного признака от своей средней на единицу и при постоянстве (фиксированном уровне) других факторов.

Значимость коэффициентов множественной регрессии оценивается на основе t-критерия Стьюдента; t_p рассчитывают как отношение взятого по модулю коэффициента регрессии к его средней ошибке с заданными уровнем значимости (α) и числом степеней свободы $d.f. = n - m - 1$.

Коэффициенты регрессии можно преобразовать в сравнимые относительные показатели - *стандартизованные коэффициенты регрессии*, или β -коэффициенты (2.5). β -коэффициент позволяет оценить меру влияния вариации факторного признака на вариацию результата при фиксированном уровне других факторов:

$$\beta_{x_i} = a_i \frac{\sigma_{x_i}}{\sigma_y}, \quad (2.5)$$

где σ_{x_i} – среднее квадратическое отклонение факторного признака,

σ_y – среднее квадратическое отклонение результативного признака,

a_i – коэффициент регрессии при соответствующем факторном признаке x_i .

При интерпретации результатов корреляционно-регрессионного анализа часто используют *частные коэффициенты эластичности* (E_{x_i}). Коэффициент эластичности (2.6) показывает, на сколько процентов в среднем изменится значение результативного признака при изменении факторного на 1% и при постоянстве (фиксированном уровне) других факторов:

$$E_{x_i} = a_i \frac{\bar{x}_i}{\bar{y}}, \quad (2.6)$$

где \bar{x}_i – среднее значение факторного признака,

\bar{y} – среднее значение результативного признака.

Множественная корреляция характеризует тесноту и направленность связи между результативным и несколькими факторными признаками. Основой измерения связей является матрица парных коэффициентов корреляции (см. п.3.2). По ней можно в первом приближении судить о тесноте связи факторных признаков между собой и с результативным признаком, а также осуществлять предварительный отбор факторов для включения их в уравнение регрессии. При этом не следует включать в модель факторы, слабо коррелирующие с результативным признаком и тесно связанные между собой. Не допускается включать в модель функционально связанные между собой факторные признаки, так как это приводит к неопределенности решения.

Более точную характеристику тесноты зависимости дают *частные коэффициенты корреляции*. Их удобно анализировать, если они представлены в табличном виде. Частный коэффициент корреляции служит показателем линейной связи между двумя признаками, исключая влияние всех остальных представленных в модели факторов. Например, для двухфакторной модели частный коэффициент корреляции между y и x_1 при фиксированном x_2 ($r_{yx1/x2}$) определяется в соответствии с (2.7).

$$r_{yx1/x2} = \frac{r_{yx1} - r_{x1x2}r_{yx2}}{\sqrt{(1 - r_{x1x2}^2)(1 - r_{yx2}^2)}}, \quad (2.7)$$

где r_{yx1} , r_{yx2} , r_{x1x2} – парные коэффициенты корреляции.

Проверка значимости частных коэффициентов корреляции аналогична, как и для парных коэффициентов корреляции.

Множественный коэффициент корреляции (R) рассчитывается при наличии линейной связи между всеми признаками регрессионной модели. R изменяется в пределах от 0 до 1. Значимость множественного коэффициента корреляции проверяется на основе F -критерия Фишера. Например, в двухфакторной модели при оценке связи между результативным и факторными признаками для определения множественного коэффициента корреляции можно использовать формулу (2.8):

$$R_{yx1x2} = \sqrt{\frac{\delta_{yx1x2}^2}{\sigma_y^2}},$$

или

$$R_{yx1x2} = \sqrt{\frac{r_{yx1}^2 + r_{yx2}^2 - 2r_{yx1}r_{yx2}r_{x1x2}}{1 - r_{x1x2}^2}}, \quad (2.8)$$

где δ_{yx1x2}^2 – дисперсия результативного признака, рассчитанная по регрессионному уравнению,

σ_y^2 – общая дисперсия результативного признака,

r_{yx1} , r_{yx2} , r_{x1x2} – парные коэффициенты корреляции.

Квадрат множественного коэффициента корреляции называют *множественным коэффициентом детерминации* (R^2). R^2 оценивает долю вариации результативного фактора за счет представленных в модели факторов в общей вариации результата. Множественный коэффициент детерминации обычно корректируют на потерю степеней свободы вариации по формуле (2.9):

$$R^2_{\text{корр}} = 1 - (1 - R^2) \frac{n-1}{n-m-1}, \quad (2.9)$$

где $R^2_{\text{корр}}$ – скорректированный множественный коэффициент детерминации,

R^2 – множественный коэффициент детерминации,

n – объем совокупности,

m – количество факторных признаков.

Статистическая надежность регрессионного уравнения в целом оценивается на основе F-критерия Фишера: проверяется нулевая гипотеза о несоответствии представленных регрессионным уравнением связей реально существующим ($H_0: a_0 = a_1 = a_2 = \dots = a_m = 0, R = 0$). Для проверки H_0 следует рассчитать значение F-критерия (F_p) и сравнить его с табличным значением (F_r), определяемым с использованием таблицы приложения 1 по заданному уровню значимости ($\alpha = 0,05$) и числу степеней свободы ($d.f.1 = m-1$ и $d.f.2 = n-m$). F_p определяется из соотношения факторной и остаточной дисперсий, рассчитанных на одну степень свободы по формуле (2.10):

$$F_p = \frac{D_{\text{факт}}}{m-1} : \frac{D_{\text{ост}}}{n-m}, \quad (2.10)$$

где $D_{\text{факт}}$, $D_{\text{ост}}$ – суммы квадратов отклонений, характеризующие факторную и остаточную вариации результативного признака. В случае однофакторного дисперсионного комплекса $D_{\text{факт}}$ и $D_{\text{ост}}$ выражаются в соответствии с (2.11),

$d.f.1 = m-1$ – число степеней свободы факторной дисперсии,

$d.f.2 = n-m$ – число степеней свободы остаточной дисперсии.

$$D_{\text{факт}} = \sum_{j=1}^m (\bar{y}_j - \bar{y})^2 n_j, \quad (2.11)$$

$$D_{\text{ост}} = \sum_{j=1}^m \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2,$$

где y_{ij} – значения результативного признака у i -й единицы в j -й группе,

i – номер единицы совокупности,

j – номер группы,

n_j – численность j -й группы,

\bar{y}_j – средняя величина результативного признака в j -й группе,
 \bar{y} – общая средняя результативного признака.

Если $F_p > F_T$, то гипотеза H_0 отвергается. При этом с вероятностью $1-\alpha = 0,95$, или 95%, принимается альтернативная гипотеза о неслучайной природе оцениваемых характеристик, т.е. признается статистическая значимость регрессионного уравнения и его параметров.

3. РЕШЕНИЕ ЗАДАЧ КОРРЕЛЯЦИОННО-РЕГРЕССИОННОГО АНАЛИЗА СТАТИСТИЧЕСКИХ СВЯЗЕЙ ПРИЗНАКОВ НА ПЕРСОНАЛЬНОМ КОМПЬЮТЕРЕ В СРЕДЕ ПАКЕТА *STATISTICA*

3.1. Общие сведения об интегрированном статистическом пакете общего назначения *STATISTICA*

В настоящем разделе дано краткое описание системы *STATISTICA*, более подробные сведения о пакете приведены в [3, 4], а также в поставляемой вместе с системой документацией фирмы-разработчика StatSoft и кратком руководстве. Следует отметить, что в процессе работы в среде *STATISTICA* студент может воспользоваться экранным справочником, содержащим практически всю информацию печатной документации. *STATISTICA* полностью удовлетворяет основным стандартам среды *Windows*:

- стандартам пользовательского интерфейса;
- технологии *DDE* — динамического обмена данными из других приложений. Благодаря поддержке DDE нетрудно выполнить командные сценарии изнутри других приложений. Например, можно в Excel написать минипрограмму (макрос), которая запускает пакет *STATISTICA*. После добавления в макрос специальных SQL-команд можно импортировать в пакет данные;
- технологии OLE — связывания и внедрения объектов, поддержка основных операций с буфером обмена и др. Использование OLE технологии обмена между Windows-приложениями позволяет легко интегрировать результаты, например, между WinWord и *STATISTICA*.

Статистический анализ данных в системе *STATISTICA* можно представить в виде следующих основных этапов:

- ввод данных в электронную таблицу с исходными данными и их предварительное преобразование перед анализом (структурирование, построение необходимых выборок, ранжирование и т. д.);
- визуализация данных при помощи того или иного типа графиков;

- применение конкретной процедуры статистической обработки;
- вывод результатов анализа в виде графиков и электронных таблиц с численной и текстовой информацией;
- подготовка и печать отчета;
- автоматизация процессов обработки при помощи макрокоманд, языка *SCL* или *STATISTICA BASIC*.

Интегрированный статистический пакет общего назначения *STATISTICA* состоит из следующих основных компонент:

- многофункциональной системы для работы с данными, которая включает в себя электронные таблицы для ввода и задания исходных данных, а также специальные таблицы (*Scroolsheet*™) для вывода численных результатов анализа. Для сложной обработки данных в *STATISTICA* имеется модуль *Управления данными*;
- графической системы для визуализации данных и результатов статистического анализа;
- набора статистических модулей, в которых собраны группы логически связанных между собой статистических процедур (рис.2):
 - основные статистики и таблицы;
 - непараметрическая статистика;
 - дисперсионный анализ;
 - множественная регрессия;
 - нелинейное оценивание;
 - анализ временных рядов и прогнозирование;
 - кластерный анализ;
 - управление данными;
 - факторный анализ и др.

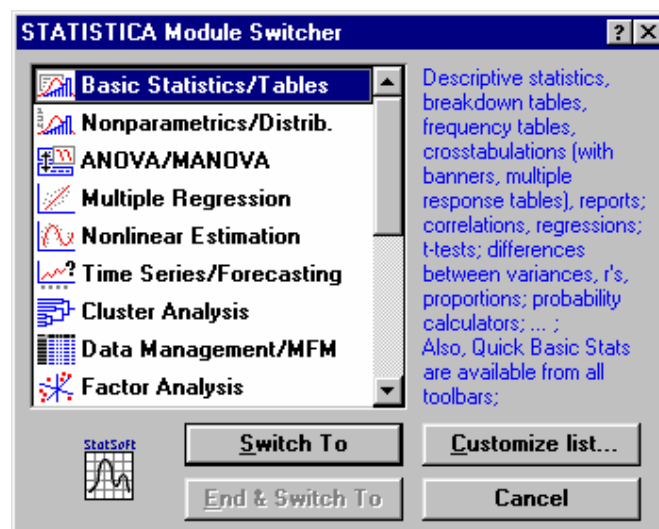


Рис. 2. Основное меню системы STATISTICA. ПЕРЕКЛЮЧАТЕЛЬ МОДУЛЕЙ

После запуска системы *STATISTICA* на экране появляется **Переключатель модулей** (рис. 2). Модули взаимодействуют друг с другом, имея одинаковый формат системных файлов. Если пользователю нужен, например, раздел линейной регрессии, то следует выбрать модуль **Multiple Regression - Множественной регрессии** и выполнить команду **Switch To**. В любом конкретном модуле можно выполнить определенный способ статистической обработки, не обращаясь к процедурам из других модулей. Все основные операции при работе с данными и графические возможности доступны в любом статистическом модуле и на любом шаге анализа. Специального инструментария для подготовки отчетов. При помощи текстового редактора, встроенного в систему, можно готовить полноценные отчеты. В пакете *STATISTICA* также имеется возможность автоматического создания отчетов;

- встроенных языков SCL и *STATISTICA BASIC*, которые позволяют автоматизировать рутинные процессы обработки данных в системе.

Способы взаимодействия с системой

Статистический анализ данных можно осуществлять в одном из следующих режимов.

Интерактивный режим работы предусматривает взаимодействие с системой при помощи последовательного выбора различных команд из меню. Этот режим предпочтителен на этапе выбора математической модели явления и метода статистического анализа. После предварительного анализа данных следует использовать другие режимы.

Использование *макрокоманд* позволяет записывать последовательность команд в одну макрокоманду. При этом можно записывать как последовательности нажатий клавиш на клавиатуре, так и движения мыши. Это удобное средство, автоматизирующее выполнение часто повторяющихся шагов статистического анализа.

Командный язык системы *STATISTICA* (язык SCL — *STATISTICA Command Language*) позволяет выполнять статистическую обработку данных в *пакетном* режиме. При этом можно установить соответствие между программой, написанной на SCL, и ярлыком в рабочем пространстве *Windows* и запускать ее как обычное *Windows*-приложение.

Язык *STATISTICA BASIC* предоставляет возможность пользователю писать собственные процедуры обработки данных.

Ввод данных

Данные в *STATISTICA* организованы в виде электронной таблицы — *Spreadsheet*. Они могут содержать как числовую, так и текстовую

информацию. Данные в электронной таблице могут иметь различные форматы, например, даты, времени и др. Электронные таблицы в *STATISTICA* поддерживают различные типы операций с данными - такие, как: операции с использованием *буфера обмена Windows*, операции с выделенными блоками значений (аналогично *MS® Excel®*), в том числе и с использованием метода *Drag-and-Drop* — «Перетащить и опустить», автозаполнение блоков и т. д. Ввести данные в электронную таблицу можно одним из следующих способов.

Непосредственно ввести их в электронную таблицу с клавиатуры. В *STATISTICA* имеются развитые инструментальные средства для автоматизации ручного ввода данных (рис. 4).

Вычислить новые данные на основе уже введенных при помощи формул, которые можно задать в электронной таблице. При этом имеется возможность быстрого доступа к большому количеству специализированных математических, статистических функций и логических операторов. Для задания сложных процедур преобразования данных можно воспользоваться встроенным языком *STATISTICA BASIC*.

Воспользоваться данными, подготовленными в другом приложении. При этом доступны следующие способы ввода данных из других приложений в систему *STATISTICA*:

- операции копирования данных через *Буфер обмена — Clipboard Windows*;
- импорт данных из наиболее популярных;
- использование механизма динамической связи *DDE* между данными в *STATISTICA* и другим *Windows*-приложением.

Для более сложных процедур обработки исходных данных в *STATISTICA* существует специализированный модуль **Data Managment** — *УПРАВЛЕНИЕ ДАННЫМИ* (рис. 2), который содержит большое количество вспомогательных процедур по работе с данными (иерархическая сортировка, проверка, ранжирование и др.)

Вывод результатов анализа

Вывести результаты анализа можно одним из следующих способов.

Численные результаты статистического анализа в системе *STATISTICA* выводятся в виде специальных электронных таблиц, которые называются таблицами вывода результатов — *Scrollsheets*™. Таблицы *Scrollsheet* могут содержать как числовую, так и текстовую информацию. Обычно даже в результате простейшего статистического анализа выдается большое количество числовой и графической информации. В системе *STATISTICA* эта информация выводится в виде последовательности, которая состоит из набора таблиц *Scrollsheet* и графиков.

STATISTICA содержит инструменты для удобного просмотра результатов статистического анализа и их визуализации. Они включают

в себя стандартные операции по редактированию таблицы (включая операции над блоками значений, *Drag-and-Drop* — «Перетащить и опустить», автозаполнение блоков и др.), операции удобного просмотра (подвижные границы столбцов, разделение прокрутки в таблице и др.), доступ к основным статистическим процедурам и графическим возможностям системы *STATISTICA*. При выводе ряда результатов (например, корреляционной матрицы) *STATISTICA* отмечает значимые параметры (например, коэффициенты корреляции) красным цветом.

Если пользователю необходимо провести *детальный статистический анализ промежуточных результатов*, то можно сохранить таблицу *Scrollsheet* в формате файла данных *STATISTICA* и далее работать с ним, как с обычными данными.

Кроме вывода результатов анализа в виде отдельных окон с графиками и таблицами *Scrollsheet* в системе *STATISTICA* имеется возможность *создания отчета*, в окно которого может быть выведена вся эта информация. Отчет — это документ (в формате *RTF*), который может содержать любую текстовую или графическую информацию. В пакете *STATISTICA* имеется возможность автоматического создания отчета (автоотчета). При этом любая таблица *Scrollsheet* или график могут автоматически быть направлены в отчет через команды меню **File/Page/Output Setup** (см. рис.3).



Рис. 3. Диалоговое окно задания параметров вывода

Таким образом, система *STATISTICA* работает с следующими типами документов: электронной таблицей *Spreadsheet* (предназначенной для ввода исходных данных), электронной таблицей *Scrollsheet* (предназначенной для вывода числовых и текстовых результатов анализа), графиком (предназначенным для визуализации численной информации), отчетом (предназначенным для вывода текстовой и графической информации в формате *RTF*).

Особенности управления пакетом

К основным преимуществам управления пакетом *STATISTICA* можно отнести следующие:

Данные можно без затруднений вводить в среду пакета, легко редактировать, создавать новые переменные, выбирать отдельные наблюдения или «вырезать» подмножество данных по строкам и (или) по столбцам таблицы «объект-признак». Благодаря обширной панели инструментов для выполнения большинства задач достаточно нескольких щелчков мышью, так как практически для всех функций пакета имеются пиктограммы. В противном случае, если студент забыл задать ту или иную переменную или параметр статистического метода, пакет сделает запрос к пользователю с необходимой подсказкой.

Особенностью пакета является настройка функций под экран, открытый в данный момент времени. Так, при загрузке пакета в активном окне возникает список модулей, доступных пользователю в данный момент времени, и пользователь может самостоятельно решить, какой вид анализа необходимо выполнить. Список модулей и порядок их следования в окне могут быть определены самим студентом, что дает ему дополнительные удобства в гибкости настройки.

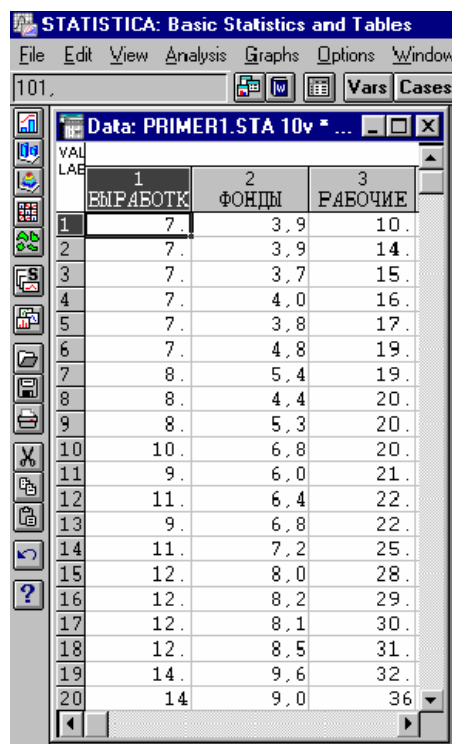
STATISTICA имеет возможность работы в пакетном режиме, используя свой командный язык *SCL*. Можно использовать и наборы команд, объединяемые в последовательности, или макросы.

Наиболее сильной стороной *STATISTICA* являются ее графические возможности. В пакете представлено множество графиков типа 2-D или 3-D, матрицы и пиктограммы. Средства управления графиками включают в себя работу одновременно с несколькими графиками, изменение размеров сложных объектов, расширенные возможности рисования и т.д.

3.2. Пример решения задачи

Условие задачи

По 20 предприятиям отрасли изучается зависимость выработки продукции на одного работника (y), тыс. руб. («ВЫРАБОТКА») от ввода в действие новых основных фондов в % от стоимости фондов на конец года (x_1) («ФОНДЫ») и от удельного веса рабочих высокой квалификации в общей численности рабочих (x_2), % - («РАБОЧИЕ»). Данные записаны в файле пакета *STATISTICA* и представлены на рис.4.



The screenshot shows the 'Data: PRIMER1.STA 10v' window in STATISTICA. It displays a table with 20 rows of data. The columns are labeled 1, 2, and 3, corresponding to 'ВЫРАБОТКА', 'ФОНДЫ', and 'РАБОЧИЕ' respectively. The data values are as follows:

	1 ВЫРАБОТКА	2 ФОНДЫ	3 РАБОЧИЕ
1	7.	3,9	10.
2	7.	3,9	14.
3	7.	3,7	15.
4	7.	4,0	16.
5	7.	3,8	17.
6	7.	4,8	19.
7	8.	5,4	19.
8	8.	4,4	20.
9	8.	5,3	20.
10	10.	6,8	20.
11	9.	6,0	21.
12	11.	6,4	22.
13	9.	6,8	22.
14	11.	7,2	25.
15	12.	8,0	28.
16	12.	8,2	29.
17	12.	8,1	30.
18	12.	8,5	31.
19	14.	9,6	32.
20	14.	9,0	36.

Рис. 4. Исходный файл с данными (Primer1.sta)

Задания

1. Получить дискриптивные статистики по каждому признаку. Оценить показатели вариации каждого признака и сделать вывод о возможностях применения метода наименьших квадратов для их изучения.
2. Составить уравнение множественной регрессии, оценить его параметры, пояснить их экономический смысл.
3. Рассчитать частные коэффициенты эластичности и дать на их основе сравнительную оценку силы влияния факторов на результат.
4. Проанализировать линейные коэффициенты парной и частной корреляции.
5. Оценить значения скорректированного и нескорректированного линейных коэффициентов множественной корреляции.

6. С помощью F-критерия Фишера оценить статистическую надежность уравнения регрессии в целом.

Решение задачи

1. Для получения дискриптивных статистик необходимо в *Переключателе модулей* (см. рис.2), появившемся после запуска пакета *STATISTICA*, выбрать команду *Basic Statistics/Tables*, при этом на экране появится стартовая панель модуля *Основные статистики и таблицы*, в которой следует выбрать команду *Descriptive statistics*. Статистическую обработку данных следует предварить открытием уже существующего файла с данными через команду *Open Data* (рис. 5) или ввести данные в компьютер через команду *File/ New Data* (рис. 4).

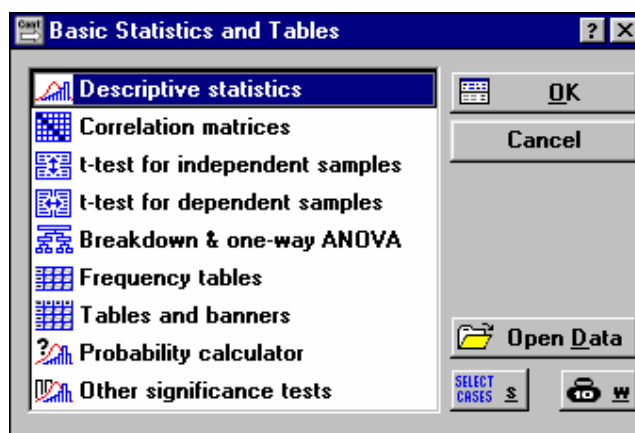


Рис. 5. Стартовая панель модуля *ОСНОВНЫЕ СТАТИСТИКИ И ТАБЛИЦЫ*

После выбора команды *OK* на экране появятся дискриптивные статистики (рис.6), анализ которых следует начать с определения показателей вариации.

STATISTICA: Basic Statistics and Tables - [Descriptive Statistics (primer1.sta)]									
File Edit View Analysis Graphs Options Window Help									
20, Columns Rows									
Continue...	Valid N	Mean	Variance	Std.Dev.	Skewness	Std.Err. Skewness	Kurtosis	Std.Err. Kurtosis	
ВЫРАБОТК	20	9,60000	6,04211	2,458069	,445096	,512103	-1,19605	,992384	
ФОНДЫ	20	6,19000	3,75884	1,938773	,188101	,512103	-1,33143	,992384	
РАБОЧИЕ	20	22,30000	46,43158	6,814072	,327801	,512103	-,53653	,992384	

Рис.6. Результаты работы модуля *ДИСКРИПТИВНЫЕ СТАТИСТИКИ*

Сравнивая значения средних величин (графа **Mean**, рис. 6), средних квадратических отклонений (графа **Standard deviation**, рис. 6), определяя коэффициент вариации ($V_y = 25,6 \%$, $V_{x_1} = 31,3 \%$, $V_{x_2} = 30,6 \%$), приходим к выводу о повышенном уровне варьирования признаков, хотя и в допустимых пределах, не превышающих 35%. Значения коэффициентов асимметрии (графа **Skewness**, рис. 6), эксцесса (графа **Kurtosis**, рис. 6) не превышают двухкратных среднеквадратических ошибок (графы **Standard error of skewness**, **Standard error of kurtosis**, рис. 6). Это указывает на отсутствие значимой скошенности и остро-(плоско)вершинности фактического распределения предприятий по значениям каждого признака по сравнению с их нормальным распределением. Совокупность предприятий однородна, и для ее изучения могут использоваться метод наименьших квадратов и вероятностные методы оценки статистических гипотез. Для построения уравнения множественной регрессии необходимо в ПЕРЕКЛЮЧАТЕЛЕ МОДУЛЕЙ (рис.1) выбрать команду **Multiple Regression**. При этом на экране появится стартовая панель модуля МНОЖЕСТВЕННАЯ РЕГРЕССИЯ (рис.7).

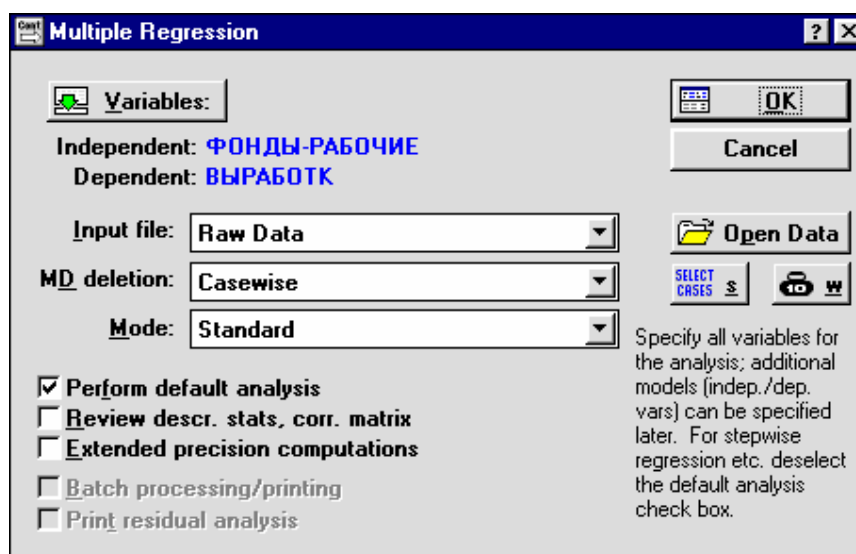


Рис.7. Стартовая панель модуля *МНОЖЕСТВЕННАЯ РЕГРЕССИЯ*

После выбора команды **Variable** (рис.7) следует указать зависимую (**ВЫРАБОТКА**) и независимые переменные (**ФОНДЫ, РАБОЧИЕ**). Выбрав команду **ОК**, получаем результаты работы модуля *МНОЖЕСТВЕННАЯ РЕГРЕССИЯ* (рис.8-9), на основе которых студент строит уравнение линейной множественной регрессии. Свободный член и коэффициенты

регрессии представлены в графе **B** (рис.8): $a_0 = 1,835$; $a_1 = 0,946$; $a_2 = 0,086$. При этом уравнение множественной регрессии примет вид: $y = 1,835 + 0,946x_1 + 0,086x_2$.

	BETA	St. Err. of BETA	B	St. Err. of B	t(17)	p-level
Intercept			1.835307	.471065	3.896080	.001162
X1	.746105	.167667	.945948	.212576	4.449917	.000351
X2	.237343	.167667	.085618	.060483	1.415561	.174964

Рис. 8. Результаты построения линейной регрессионной модели

Для оценки значимости полученных коэффициентов регрессионного уравнения воспользуемся t-критерием Стьюдента (графа **t(17)**, рис. 8). В пакете *STATISTICA* значения t-критерия (t_p) определяются как отношение взятого по модулю коэффициента регрессии (графа **B**, рис. 8) к его стандартной ошибке (графа **St. Err. of B**, рис. 8). Табличное значение t-критерия с уровнем значимости $\alpha=0,01$ и числом степеней свободы $d.f.=n-m-1=17$: $t_r = 2,89$ (прил.2). Сравним значения t_p и t_r для каждого из полученных параметров:

- $t_p = 3,89 > t_r$ - для свободного члена a_0 ;
- $t_p = 4,44 > t_r$ - для коэффициента a_1 ;
- $t_p = 1,41 < t_r$ - для коэффициента a_2 .

Таким образом, статистически значимыми являются коэффициенты a_0 и a_1 , а коэффициент a_2 сформирован под влиянием случайных причин. Поэтому фактор x_2 можно исключить из модели как неинформативный. Аналогичный вывод можно сделать, сравнивая значения уровня значимости (графа **p-level**, рис. 8) с принятым нами уровнем $\alpha=0,01$. Для a_0 и a_1 показатель вероятности случайных значений параметров регрессии меньше 1% ($0,01 \cdot 100\%$). Поэтому справедлив вывод о том, что полученные коэффициенты статистически значимы и надежны. Для a_2 делается вывод о случайной природе его значения, поскольку $\alpha = 0,175 \cdot 100\% = 17,5\% > 1\%$. Это позволяет рассматривать x_2 как неинформативный фактор. Его можно удалить из уравнения для улучшения модели.

Свободный член a_0 оценивает агрегированное влияние прочих (кроме учтенных в модели x_1 и x_2) факторов на результат y . Коэффициенты a_1 и a_2 указывают на то, что с увеличением x_1 и x_2 на единицу их значений y увеличивается соответственно на 0,9459 тыс.руб. и на 0,0856 тыс.руб. Сравнить эти значения не следует, так как они зависят от единиц измерения каждого признака и потому несопоставимы между собой. Для сравнения можно воспользоваться сравнимыми относительными показателями - β -коэффициентами (графа **BETA**, рис. 8).

3. Для определения частных коэффициентов эластичности в соответствии с (2.6) воспользуемся коэффициентами регрессионного уравнения a_1 и a_2 и значениями средних величин результативного и факторных признаков (графа **Mean**, рис.6). $E_{x_1} = 0,61\%$, $E_{x_2} = 0,19\%$. Полученные коэффициенты показывают, что с увеличением коэффициента обновления основных фондов (x_1) на 1% от его среднего уровня выработка продукции на одного работника (y) увеличится на 0,61%, от своего среднего уровня. Аналогично с увеличением доли рабочих высокой квалификации в общей численности рабочих (x_2) на 1% от ее среднего уровня выработка продукции на одного работника (y) увеличится на 0,19%, от своего среднего уровня. По значениям частных коэффициентов эластичности можно сделать вывод о более сильном влиянии на результат фактора обновления основных фондов, чем доли рабочих высокой квалификации. Для построения парных коэффициентов корреляции включенных в модель факторов можно через матрицу парных коэффициентов корреляции, а тесноту связи значений двух переменных, исключая влияние всех других переменных, представленных в уравнении множественной регрессии, можно через матрицу линейных коэффициентов частной корреляции. Для построения этих матриц в модуле *МНОЖЕСТВЕННАЯ РЕГРЕССИЯ* (рис.9) следует последовательно выбрать команды ***Correlations and desc.stats*** (для построения матрицы парных коэффициентов корреляции), ***Partial correlations*** (для построения матрицы линейных коэффициентов частной корреляции).

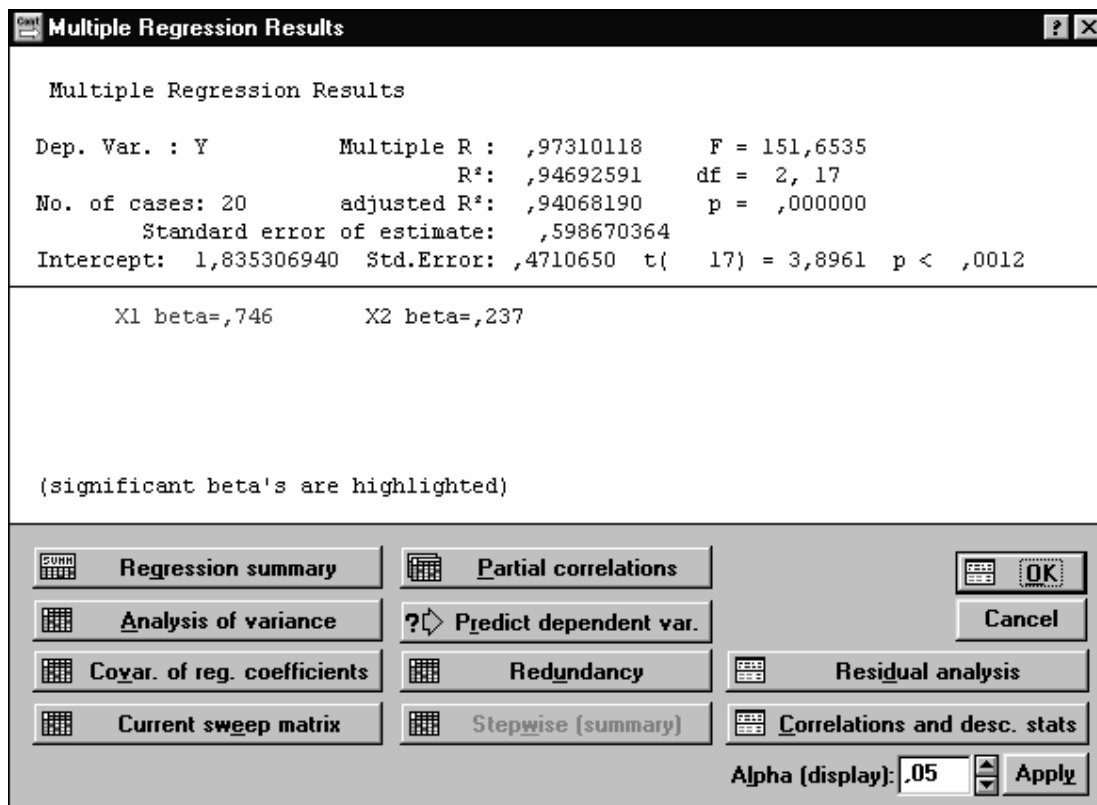


Рис.9. Результаты построения линейной регрессионной модели

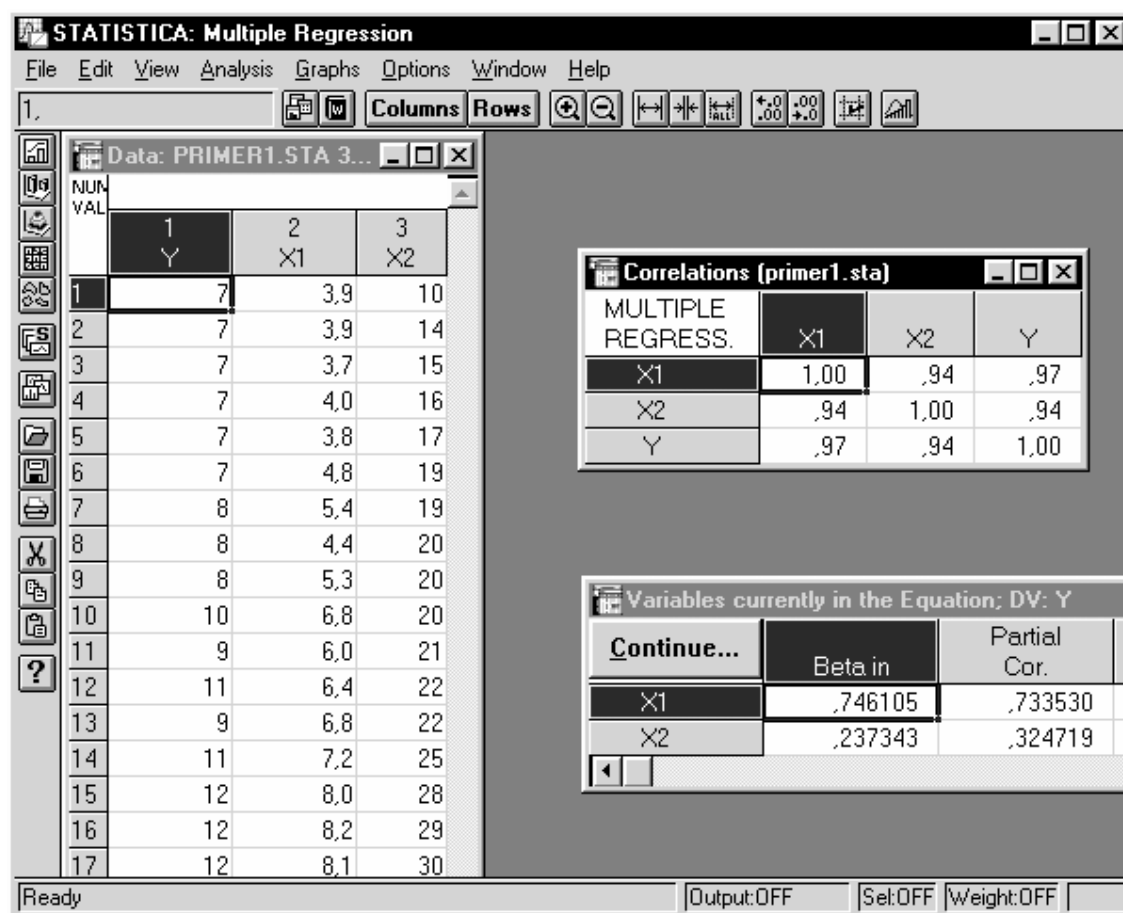


Рис. 10. Результаты построения корреляционных матриц

Полученные значения парных коэффициентов корреляции говорят о тесной связи выработки продукции на одного работника (y) как с коэффициентом обновления основных фондов (x_1) - $r_{yx1} = 0,97$, так и с долей рабочих высокой квалификации в общей численности рабочих (x_2) - $r_{yx2} = 0,94$. При этом следует учитывать тесную межфакторную связь x_1 с x_2 ($r_{x1x2} = 0,94$), примерно равную связи y с x_2 . Поэтому для улучшения модели фактор x_2 можно исключить как недостаточно статистически надежный. Коэффициенты частной корреляции дают более точную характеристику тесноты зависимости двух признаков, чем коэффициенты парной корреляции, так как «очищают» парную зависимость от взаимодействия данной пары признаков с другими признаками, представленными в модели. Наиболее тесно показатель выработки продукции на одного работника (y) связан с коэффициентом обновления основных фондов (x_1) - $r_{yx1/x2} = 0,73$ - по сравнению со связью y с долей

рабочих высокой квалификации в общей численности рабочих (x_2) - $r_{yx2/x1} = 0,32$. Этот факт также говорит в пользу исключения фактора x_2 из модели.

5. Коэффициенты линейной множественной корреляции (детерминации) представлены на рис. 8-9. Коэффициент множественной корреляции $R_{yx1x2} = 0,973$ свидетельствует о тесной связи факторных признаков с результативным.

Нескорректированный коэффициент множественной детерминации $R^2_{yx1x2} = 0,947$ оценивает долю вариации результата за счет представленных в уравнении факторов в общей вариации результата. Он указывает на высокую степень обусловленности вариации результата вариацией факторных признаков. Скорректированный коэффициент множественной детерминации $R^2_{yx1x2} = 0,941$ оценивает тесноту связи с учетом степеней свободы (см. п.2.2), что позволяет его использовать для оценки тесноты связи в моделях с разным числом факторов.

Значения коэффициентов множественной детерминации позволяют сделать вывод о высокой (более 90%) детерминированности результативного признака y в модели факторными признаками x_1 и x_2 .

6. Оценим статистическую надежность полученного уравнения множественной регрессии с помощью общего F-критерия, который проверяет нулевую гипотезу о статистической незначимости параметров построенного регрессионного уравнения и показателя тесноты связи ($H_0: a_0 = a_1 = a_2 = 0, R_{yx1x2} = 0$).

Фактическое значение F-критерия Фишера - $F_p = 151,7$ (см. рис. 8-9). Сравним его с табличным значением F-критерия, определяемым с использованием таблицы приложения 1 по заданным уровню значимости ($\alpha = 0,05$) и числу степеней свободы (в пакете *STATISTICA* $d.f.1 = m = 2$ и $d.f.2 = n - m - 1 = 17$). $F_t = 3,59$. Поскольку $F_p > F_t$, то гипотеза H_0 отвергается. Так как вероятность случайного значения F_p значительно меньше 5% ($p < 0,000001$, см. рис. 8-9), то с вероятностью более 95% принимается альтернативная гипотеза. Таким образом, признается статистическая значимость регрессионного уравнения, его параметров и показателя тесноты связи R_{yx1x2} .

3.3. Порядок выполнения индивидуального задания

1. *Ввод исходных данных.* Получив индивидуальное задание, студент создает файл с именем *.sta и заносит в него данные. Файл следует сохранить в указанном преподавателем каталоге.

2. *Дикриптивно-статистический анализ данных.* На данном этапе выполнения работы определяются значения средних величин, средних квадратических отклонений, значения коэффициентов асимметрии, эксцесса и их среднеквадратических ошибок по результативному и факторным признакам. Студенту следует оценить показатели вариации каждого признака и сделать вывод о возможностях применения метода наименьших квадратов для их изучения, а если необходимо, то исключить резко отклоняющиеся единицы совокупности.

3. *Построение уравнения множественной регрессии.* На этом этапе определяются коэффициенты множественной регрессии, составляется регрессионное уравнение, оцениваются его параметры.

4. *Определение частных коэффициентов эластичности.* Студент самостоятельно рассчитывает частные коэффициенты эластичности и дает на их основе сравнительную оценку силы влияния факторов на результат.

5. *Анализ линейных коэффициентов парной и частной корреляции.* Данный этап предусматривает построение матриц коэффициентов парной и частной корреляции и оценку целесообразности включения факторных признаков в модель.

6. *Оценка коэффициентов множественной корреляции (детерминации).*

7. *Оценка статистической надежности полученного уравнения регрессии.*

8. *Оформление отчета.* Титульный лист отчета должен содержать название работы, цель работы, фамилию, инициалы, курс и группу студента, выполнившего индивидуальное задание. В отчете следует отразить основные этапы выполненного задания, полученные результаты и сделать выводы по каждому этапу. Для этой цели можно использовать распечатки отчета, полученного средствами пакета *STATISTICA* (файл с расширением *.rtf), включая его широкие графические возможности.

9. *Защита индивидуального задания.* Защита индивидуального задания преследует цель оценить знания студента по вопросам построения регрессионных моделей с помощью СПП *STATISTICA* и интерпретации результатов корреляционно-регрессионного анализа данных. При подготовке к защите индивидуального задания студенту следует ответить на представленные в п.4 вопросы.

4. ВОПРОСЫ ДЛЯ САМОПРОВЕРКИ

1. Дайте определение функциональному, статистическому и корреляционному типам связи.

2. Назовите основные условия применения корреляционно-регрессионного метода анализа статистических связей.
3. Для решения каких типов задач используется корреляционно-регрессионный метод?
4. Приведите примеры различных видов уравнений парной и множественной регрессии.
5. Дайте определение парному и множественному линейным коэффициентам корреляции.
6. Как оценивается значимость коэффициента корреляции?
7. Чем характеризуются функционально связанные между собой факторы?
8. Что характеризуют параметры регрессионного уравнения? Объясните сущность коэффициента парной линейной регрессии.
9. В чем заключается метод наименьших квадратов? Каковы основные условия его применения?
10. Как оценивается значимость параметров регрессионного уравнения?
11. Дайте определение частному коэффициенту эластичности. Что он характеризует?
12. Дайте определение стандартизованному коэффициенту регрессии. Что он характеризует?
13. Что позволяет оценить множественный коэффициент детерминации?
14. Для чего используется скорректированный множественный коэффициент детерминации?
15. Как оценить статистическую надежность регрессионного уравнения в целом?

СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ

1. Айвазян С.А. Программное обеспечение персональных ЭВМ по статистическому анализу данных // Компьютер и экономика: экономические проблемы компьютеризации общества. М.: Наука, 1991. С. 91–107.
2. Айвазян С.А., Степанов В.С. Инструменты статистического анализа данных // «Мир ПК». 1997. №8. С. 33–41.
3. Боровиков В.П., Боровиков И.П. STATISTICA. Статистический анализ и обработка данных в среде Windows. М.: Филин, 1997.
4. Боровиков В.П. Популярное введение в программу STATISTICA. М., 1998.
5. Векслер Л.С. Статистический анализ на персональном компьютере // «Мир ПК». 1992. №2. С. 89–97.

6. Елисеева И.И., Юзбашев М.М. Общая теория статистики. М.: Финансы и статистика, 1998.
7. Ефимова М.Р., Петров Е.В., Румянцев В.Н. и др. Общая теория статистики / Под ред. проф. М.Р. Ефимовой. М.: ИНФРА-М, 1998.
8. Костеева Т.В., Курышева С.В., Михайлов Б.А. Эконометрика: Решение типовых задач. СПб.: СПбГУЭФ, 1997.
9. Крастинь О.П. Разработка и интерпретация моделей корреляционных связей в экономике. Рига: Зинате, 1983.
10. Теория статистики: Учебник / Под ред. проф. Р.А. Шмойловой. М.: Финансы и статистика, 1998.
11. Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере. М.: ИНФРА-М, 1998.

Приложение 1

Значения F-критерия Фишера при уровне значимости $\alpha=0,05$

$d.f._1$ $d.f._2$	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	254,32
2	18,5	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67
28	4,20	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
35	4,12	3,26	2,87	2,64	2,48	2,37	2,22	2,04	1,83	1,57
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,52
45	4,06	3,21	2,81	2,58	2,42	2,31	2,15	1,97	1,76	1,48
50	4,03	3,18	2,79	2,56	2,40	2,29	2,13	1,95	1,74	1,44
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
70	3,98	3,13	2,74	2,50	2,35	2,23	2,07	1,89	1,67	1,35
80	3,96	3,11	2,72	2,49	2,33	2,21	2,06	1,88	1,65	1,31

Окончание табл.

$d.f._1$ $d.f._2$	1	2	3	4	5	6	8	12	24	∞
90	3,95	3,10	2,71	2,47	2,32	2,20	2,04	1,86	1,64	1,28
100	3,94	3,09	2,70	2,46	2,30	2,19	2,03	1,85	1,63	1,26
125	3,92	3,07	2,68	2,44	2,29	2,17	2,01	1,83	1,60	1,21
150	3,90	3,06	2,66	2,43	2,27	2,16	2,00	1,82	1,59	1,18
200	3,89	3,04	2,65	2,42	2,26	2,14	1,98	1,80	1,57	1,14
300	3,87	3,03	2,64	2,41	2,25	2,13	1,97	1,79	1,55	1,10
400	3,86	3,02	2,63	2,40	2,24	2,12	1,96	1,78	1,54	1,07
500	3,86	3,01	2,62	2,39	2,23	2,11	1,96	1,77	1,54	1,06
1000	3,85	3,00	2,61	2,38	2,22	2,10	1,95	1,76	1,53	1,03
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	

Приложение 2

Значения t-критерия Стьюдента при уровне
значимости 0,10; 0,05; 0,01 (двухсторонний)

Число степеней свободы $d.f.$	α		
	0,10	0,05	0,01
1	6,3138	12,706	63,657
2	2,9200	4,3027	9,9248
3	2,3534	3,1825	5,8409
4	2,1318	2,7764	4,6041
5	2,0150	2,5706	4,0321
6	1,9432	2,4469	3,7074
7	1,8946	2,3646	3,4995
8	1,8595	2,3060	3,3554
9	1,8331	2,2622	3,2498
10	1,8125	2,2281	3,1693
11	1,7959	2,2010	3,1058
12	1,7823	2,1788	3,0545
13	1,7709	2,1604	3,0123
14	1,7613	2,1448	2,9768
15	1,7530	2,1315	2,9467
16	1,7459	2,1199	2,9208
17	1,7396	2,1098	2,8982
18	1,7341	2,1009	2,8784
19	1,7291	2,0930	2,8609
20	1,7247	2,0860	2,8453
21	1,7207	2,0796	2,8314
22	1,7171	2,0739	2,8188

Окончание табл.

Число степеней свободы $d.f.$	α		
	0,10	0,05	0,01
23	1,7139	2,0687	2,8073
24	1,7109	2,0639	2,7969
25	1,7081	2,0595	2,7874
26	1,7056	2,0555	2,7787
27	1,7033	2,0518	2,7707
28	1,7011	2,0484	2,7633
29	1,6991	2,0452	2,7564
30	1,6973	2,0423	2,7500
40	1,6839	2,0211	2,7045
60	1,6707	2,0003	2,6603
120	1,6577	1,9799	2,6174
∞	1,6449	1,9600	2,5758