**Содержание**

Введение

1. Теоретические сведения

1.1 Статистика. Виды статистического анализа

1.2 Статистический пакет STATISTICA

2. Статистический анализ экономических данных в STATISTICA

2.1 Практическое задание. Корреляционно-регрессионный анализ в STATISTICA

2.2 Практическое задание

2. Кластерный анализ в STATISTICA

Введение

Для обеспечения эффективности производства необходимо четко знать:

* Как анализировать и прогнозировать процессы?
* Как повысить эффективность производства и сократить затраты?
* Как обеспечить стабильность показателей качества продукции?
* Как принять оптимальное решение на основе анализа данных?
* Как организовать управление корпоративными источниками информации?

Для решения этих и многих других задач в области социологических и маркетинговых исследований, прогнозирования и управления качеством необходимы знания математической и экономической статистики. Принятие любого технического, финансового и производственного решения немыслимо без статистического анализа информации; выделять закономерности из случайностей, сравнивать вероятные альтернативы выбора, строить прогнозы развития процессов, обнаруживать связи и различия множества объектов возможно только и исключительно средствами математической статистики.

Статистика позволяет адекватно оценивать складывающуюся ситуацию и выявлять тенденции, принимать оперативные и стратегические решения. В условиях современной рыночной экономики статистическая информация стала важным инструментом борьбы и выживания на рынке. Поэтому пакеты статистического анализа данных являются настольным рабочим инструментом специалистов любого уровня. А для специалиста в области управления и экономики знание статистических методов обработки информации и современных компьютерных технологий, которые позволят автоматизировать громоздкие расчеты, абсолютно необходимы.

Современный экономист должен владеть несколькими основными программными средствами, в которых заложены методы статистического анализа. MS Excel предлагает широкий диапазон средств для анализа статистических и экспериментальных данных. В группу статистических функций входят функции корреляционного анализа. Кроме встроенных средств можно использовать надстройку Пакет анализа (Сервис/Надстройки/Пакет Анализа) для проведения регрессионного и дисперсионного анализа.

MathCad также имеет развитый аппарат работы с задачами математической статистики и обработки эксперимента. Во-первых, имеется большое количество встроенных специальных функций, позволяющих рассчитывать плотности вероятности и другие основные характеристики основных законов распределения случайных величин. Наряду с этим, в MathCad запрограммировано соответствующее количество генераторов псевдослучайных чисел для каждого закона распределения, что позволяет эффективно проводить моделирование методами Монте-Карло. Во-вторых, предусмотрена возможность построения гистограмм и расчета статистических характеристик выборок случайных чисел и случайных процессов, таких как средние, дисперсии, корреляции и т. п. При этом случайные последовательности могут как создаваться генераторами случайных чисел, так и вводиться из файлов. В-третьих, имеется целый арсенал средств, направленных на интерполяцию-экстраполяцию данных, построение регрессии по методу наименьших квадратов, фильтрацию сигналов. Наконец, реализован ряд численных алгоритмов, осуществляющих расчет различных интегральных преобразований, что позволяет организовать спектральный анализ различного типа.

Однако ведущим пакетом статистического анализа является система STATISTICA, которая основана на самых современных технологиях, полностью соответствует последним достижениям в области IT, позволяет решать любые задачи в области анализа и обработки данных, идеально подходит для применения в любой области: маркетинге, финансах, страховании, экономике, бизнесе, промышленности, медицине и др. Например, решение поставленных в начале статьи задач, может быть реализовано на базе промышленных модулей STATISTICA (карты контроля качества, планирование экспериментов, анализ процессов).

# 1. Теоретические сведения

# 1.1 Статистика. Виды статистического анализа

**Статистика** – наука о сборе, измерении и анализе массовых количественных данных. Статистические данные всегда являются приближенными, усредненными. Поэтому они носят оценочный характер и для достоверности результатов необходимо большое число исходных данных.

Существует несколько видов статистического анализа данных: корреляционный, регрессионный, дисперсионный, факторный, кластерный и др. Рассмотрим некоторые из них.

**Корреляционный анализ**

Иногда корреляцию и регрессию рассматривают как совокупный процесс статистического исследования. **Корреляционно-регрессионный** анализ является одним из значимых методов построения математических моделей в экономике и считается одним из главных методов в маркетинге.

Корреляция в широком смысле слова означает связь между объективно существующими явлениями.

**Корреляционный анализ** – вид статистического анализа, который состоит в количественной оценке **силы** и **направления связи** между двумя (**парная корреляция**) или несколькими (**множественная корреляция**) наборами данных. Для количественной оценки силы связи используются коэффициенты парной корреляции *r* и множественной корреляции *R*.

**Коэффициент корреляции** (безразмерная величина) – количественный показатель **линейной связи** между двумя или более наборами данных, значение которого лежит в интервале от -1 до 1. Если коэффициент равен ± 1, то связь функциональная, если равен 0, то связь отсутствует.

Для качественной оценки **силы связи** используются специальные табличные соотношения (например, шкала Чеддока, табл. 1)

Таблица 1 – Шкала Чеддока

|  |  |
| --- | --- |
| Значения коэффициента корреляции | Характер связи |
|  | Очень слабая |
|  | Слабая |
|  | Заметная |
|  | Сильная |
|  | Очень сильная |

**Направление связи** определяется знаками ±: близость к +1 означает, что возрастанию одного набора значений соответствует возрастание другого набора, близость к -1 означает обратное.

Для наглядности измерения всех связей в случае множественной корреляции целесообразно использовать **корреляционную матрицу** – матрицу из попарных коэффициентов корреляции.

**Регрессионный анализ**

**Регрессионный анализ** – вид статистического анализа, который состоит в представлении зависимости одних факторов от других в виде некоторой функции (**уравнения регрессии**) с помощью которой осуществляется прогнозирование и поиск ответа на вопросы «Что будет через какое-то время?» или «Что будет, если…?».

В случае **парной регрессии** уравнение определяется подвум наборам данных, один из которых представляет значения зависимой переменной *y*, а другой – независимой переменной *х*. В случае **множественной регрессии** уравнение определяется по нескольким наборам данных, один из которых представляет значения зависимой переменной *y*, а другие независимыми переменными *х1, х2,…,* *xm*.

Получение уравнения регрессии происходит в два этапа: подбор вида функции и вычисление параметров функции.

Выбор функции, в большинстве случаев, производятся среди линейной, квадратичной, степенной и др. видов функций (табл. 2). К функции предъявляются следующие требования: она должна быть достаточно простой для использования ее в дальнейших вычислениях и график этой функции должен проходить вблизи экспериментальных точек так, чтобы сумма квадратов отклонений *y*-координаты всех экспериментальных точек от *y*-координат графика функции была ба минимальной (метод наименьших квадратов).

Таблица 2 – Виды функций, применяемых в регрессионных моделях

|  |  |
| --- | --- |
| **Парная (простая) регрессия** | **Множественная регрессия** |
| *Линейная регрессия* |
| *y=ax+b,* | *y = а0 + a1x1+ … +amxm* |
| *Квадратичная (параболическая)* |
| *y=ax2+bx+c* | *y= а0 + a1x12 + … +am xm2* |
| *Степенная* |
| *y=axb* | *y = а0 x1 a1 x2 a2… xm am* |
| *Логарифмическая y=alnx+b,* | *Гиперболическая**y = а0 + a1 (1/x1) + … +am(1/xm)* |
| *Экспоненциальная y=aebx* |
| где *a, b, c –* коэффициенты парной регрессии. | где *а0*, *a1,a2,…,am*– коэффициенты множественной регрессии,*n* – объем совокупности,*m* – количество факторных признаков. |

**?** Какой вид регрессионного анализа (парный или множественный) в большей степени отвечает реальным условиям?

**?** Можно ли учесть все факторы *х1, х2,…*, *xm, …*в случае множественной корреляции?

Для количественной оценки точности построения уравнения регрессии предназначен **коэффициент детерминации *R2***, равный квадрату коэффициента корреляции и указывающий, какой процент изменения функции *у* объясняется воздействием факторов *хk*. Чем его значение ближе к 1, тем уравнение точнее описывает исследуемую зависимость.

Значимое уравнение (с *R2* близким к 1*)* используется, как правило, для **прогнозирования** изучаемого явления. **Прогноз** – это вероятностное суждение о будущем, полученное путем использования совокупности научных методов. Например, прогнозирование финансового состояния выполняется для того, чтобы получить ответы на два вопроса: «как это может быть (какими могут стать финансовые показатели, если не будут приняты меры по их изменению)» и «как это должно быть (какими должны стать финансовые показатели фирмы для того, чтобы ее финансовое состояние обеспечивало высокий уровень конкурентоспособности)». Прогнозирование с целью получения ответа на первый вопрос принято называть исследовательским, на второй – нормативным.

Существует два способа прогнозов по уравнению регрессии: в пределах экспериментальных значений (**интерполяция)** иза пределами (**экстраполяция**). Применимость всякой регрессионной модели ограничена, особенно за пределами экспериментальной области, т.к. характер зависимости может существенно измениться. Поэтому достоверность исследовательского прогноза может быть невысокой. Однако его выполнение полностью обосновано.

# 1.2 Статистический пакет STATISTICA

Так как статистические методы находят широкое применение во всех сферах производства, то рынок компьютерных технологий предлагает большое количество прикладных программ, которые позволяют проводить такой анализ. Обилие систем, создатели которых утверждают, что их программа является наилучшей для обработки данных, а также отсутствие у большинства специалистов достаточного времени для освоения нескольких пакетов приводит к усложнению процесса выбора. Однако, по данным statsoft.ru, лидером статистических пакетов является STATISTICA.

**История развития, области применения**

STATISTICA (американской компании StatSoft, http://www.statsoft.com, StatSoft RUSSIA – российское представительство StatSoft) – система, реализующая известные методы статистической обработки и визуализации данных, управления базами данных и разработки пользовательских приложений при помощи встроенного языка программирования Statistica Basic.

Пакет разработан в 1984 г., и первоначально он был представлен в виде модуля для самой популярной в то время электронной таблицы Lotus. Как самостоятеный продукт Statistica впервые заявила о себе в 1991 г. и с тех пор постоянно занимает лидирующее место среди специализированных пакетов по статистике.

Благодаря широкому набор процедур анализа STATISTICA применяется в научных исследованиях, технике, бизнесе. Также система хорошо зарекомендовала себя в страховании (например, в страховой компании РОСНО). STATISTICA широко используется в учебном процессе (в Московском государственном университете, например, на механико-математическом и экономическом факультетах, в Московском институте электроники и математики на экономическом факультете и факультете прикладной математики, в Московском экономико-статистическом институте и др.). Помимо общих статистических и графических средств в системе имеются специализированные модули, например, для проведения социологических или биомедицинских исследований, решения технических и, что очень важно, промышленных задач: Карты контроля качества, Анализ процессов и Планирование эксперимента. Модуль Карта контроля позволяет автоматизировать процесс контроля за качеством производимой продукции, анализировать причины появления отклонений от плановых спецификаций. Statistica осуществляет анализ пригодности (пригодности процессов/механизмов), как одной из важнейших характеристик производственного процесса. Вычисление показателей (или индексов) пригодности позволяет дать ответ на важный вопрос: какое количество изделий попадает в заданные границы инженерного допуска?

Таким образом, STATISTICA является одной из наиболее простых для неподготовленного пользователя систем, с наименьшим периодом овладевания ее возможностями и удачным набор графических возможностей.

**Интерфейс, основные возможности**

Наборы файлов данных системы STATISTICA (расширение \*.sta) можно рассматривать как “рабочие книги” файлов, поскольку они содержат и автоматически сохраняют информацию обо всех дополнительных файлах (например, графиках, отчетах и программах), которые используются с текущим набором данных.

STATISTICA использует стандартный интерфейс электронных таблиц. Текущий файл данных всегда отображается в виде электронной таблицы. Данные организованы в виде наблюдений и переменных. Наблюдения можно рассматривать как эквивалент столбцов электронной таблицы. Каждое наблюдение состоит из набора значений переменной.

Рис. 1

Система состоит из ряда модулей, работающих независимо. Каждый модуль включает определенный класс процедур. Почти все процедуры являются интерактивными, т.е. для запуска обработки необходимо выбрать из меню переменные и ответить на ряд вопросов системы. Это очень удобно для начинающего пользователя, однако резко замедляет деятельность опытного и не позволяет эффективно повторять одну и ту же процедуру несколько раз.

**Модули и процедуры**

Описательные статистики

Анализ многомерных таблиц

Подгонка распределений

Корреляционный анализ

Регрессионный анализ (в том числе и многомерный, нелинейный)

Дисперсионный анализ

Кластерный анализ

Дискриминантный анализ

Факторный анализ

Анализ соответствий

Многомерное шкалирование

Анализ выживаемости

Структурные модели

Деревья классификаций

Прогнозирование временных рядов

Непараметрическая статистика

Анализ Монте-Карло и др.

**Виды анализа**

Basic Statistics/Tables (Основные статистики и таблицы):

Advanced Linear/Nonlinear Models (прогрессивные линейные/нелинейные модели):

Multivariate Exploratory Techniques (Многомерные Исслед. Методы):

Industrial Statistic and Six Sigma (промышленная статистика и статистика 6-ти сигм):

**Графическое представление данных в STATISTICA**

STATISTICA позволяет строить различные типы графиков:

Матричные графики

ПиктографикиДиаграммы рассеяния

ГистограммыТернарные графики

Карты линий уровняКруговые диаграммы

Категоризованные

Вероятностные

Графики поверхностей

Трассировочные

Комбинированные

Вращение и перспектива

Подгонка, сглаживание, сечения и др.:

|  |  |  |
| --- | --- | --- |
| Типы графиков вSTATISTICA | Виды 2D графиков | Виды 3D графиков |
| Виды 3D линий |

#

# 2. Статистический анализ экономических данных в STATISTICA

#

# 2.1 Практическое задание 1. Корреляционно-регрессионный анализ в STATISTICA

**Постановка задачи**

Руководство компании по результатам производственной деятельности 15 своих филиалов в различных городах России анализирует факторы, влияющие на производительность труда (*y)* и предполагает, что важнейшими из них являются следующие:

*x1 –* среднегодовая стоимость основных фондов, тыс. руб.

*х2* – удельный вес рабочих высокой квалификации в общей численности рабочих, %

*х3* – трудоемкость единицы продукции

*х4* – среднегодовая численность рабочих

*x5* – коэффициент сменности оборудования

*x6* – удельный вес потерь от брака

*x7* – среднегодовой фонд заработной платы, тыс. руб.

Были собраны данные за последний год (см. таб. 3).

Таблица 3 – Исходные данные

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| № | Город | *y* | *х1* | *х2* | *х3* | *х4* | *x5* | *x6* | *x7* |
| 1 | Москва | 14 | 101,03 | 35 | 0,4 | 15780 | 2,01 | 0,22 | 13002 |
| 2 | Санкт-Петербург | 14,02 | 98,54 | 36 | 0,42 | 14760 | 1,86 | 0,25 | 10145,6 |
| 3 | Нижний-Новгород | 7,03 | 49 | 17 | 1,83 | 630 | 0,95 | 0,5 | 5040,9 |
| 4 | Ульяновск | 7,01 | 50 | 17 | 1,85 | 633 | 0,93 | 0,52 | 5027,39 |
| 5 | Пенза | 8,21 | 57,42 | 19 | 1,43 | 752 | 1,08 | 0,44 | 5903,3 |
| 6 | Самара | 10 | 70 | 24 | 1,01 | 920 | 1,33 | 0,35 | 7100 |
| 7 | Чебоксары | 9,02 | 61,03 | 22 | 1,23 | 830 | 1,19 | 0,39 | 6494,6 |
| 8 | Саранск | 11 | 78,09 | 26 | 0,82 | 1028 | 1,44 | 0,37 | 7500 |
| 9 | Челябинск | 9,05 | 63,31 | 28 | 1,2 | 804 | 1,2 | 0,38 | 6516,5 |
| 10 | Тольятти | 11 | 77,05 | 29 | 0,81 | 1028 | 1,46 | 0,32 | 7940 |
| 11 | Волгоград | 12 | 84,03 | 27 | 0,64 | 1126 | 1,6 | 0,29 | 8900 |
| 12 | Рязань | 12 | 83 | 29 | 0,66 | 1127 | 1,59 | 0,25 | 8668 |
| 13 | Красноярск | 12 | 84 | 30 | 0,68 | 1096 | 1,59 | 0,29 | 8670,91 |
| 14 | Тула | 7,26 | 50,81 | 17 | 1,75 | 657 | 0,96 | 0,49 | 5209,8 |
| 15 | Казань | 7,01 | 55,01 | 16 | 1,85 | 631 | 0,93 | 0,51 | 5027,3 |

С использованием системы STATISTICA необходимо:

1) для *y* и переменных, соответствующих варианту (см. таб. 4), построить матрицу частных коэффициентов корреляции (корреляционную матрицу). Изобразить матрицу в графическом виде.

Таблица 4 – Варианты заданий

|  |  |  |
| --- | --- | --- |
| Вариант*j* | Независимые переменные(факторные признаки) | Задания по прогнозированию |
| Как изменится производительность труда на московском предприятии, если |
| 0 | *х1, х2, х4, x5* | среднегодовую численность рабочих сократить на 780 человек, а коэффициент сменности оборудования повысить до 3? |
| 1 | *х1, х3, х4, x5* | среднегодовую стоимость основных фондов увеличить на 80 тыс. руб., а и трудоемкость единицы продукции на 0,6? |
| 2 | *х3, х4, x5, x6* | трудоемкость единицы продукции сократить в 4 раза, а коэффициент сменности оборудования снизить в 2 раза? |
| 3 | *х1, х2, х3, x5* | среднегодовую стоимость основных фондов увеличить на 60 тыс. руб., а коэффициент сменности оборудования – на 0,9? |
| 4 | *х1, х2, x6, x7* | среднегодовую стоимость основных фондов сократить до 90 тыс. руб., а удельный вес потерь от брака понизить в 2 раза? |
| 5 | *х1, х3, х4, x7* | среднегодовую стоимость основных фондов сократить до 95 тыс. руб., а трудоемкость единицы продукции понизить на 0,1? |
| 6 | *х1, х2, x5, x7* | коэффициент сменности оборудования увеличить в 2 раза, а среднегодовой фонд заработной платы уменьшить на 92 тыс. руб.? |
| 7 | *х4, x5, x6, x7* | коэффициент сменности оборудования уменьшить в 2 раза, а среднегодовой фонд заработной платы увеличить на 92 тыс. руб. |
| 8 | *х2, х3, x5, x7* | коэффициент сменности оборудования увеличить на 1,5, а среднегодовой фонд заработной платы уменьшить на 32 тыс. руб.? |
| 9 | *х1, х3, x5, x7* | коэффициент сменности оборудования уменьшить на 1,5, а среднегодовой фонд заработной платы увеличить на 32 тыс. руб.? |

2) построить линейное уравнение множественной регрессии, выбрав в качестве зависимой переменной – *y*, в качестве независимых – переменные *хi*, соответствующие варианту (см. таб. 4).

3) Определить коэффициент множественной корреляции и коэффициент детерминации *R2* полученной модели

4) Проверить значимость построенной модели (например, используя уровень значимости *α=0,05*).

5) Если модель значима дать оценку коэффициентов множественной регрессии на основе *t*-критерия, если *tтабл(15-4-1)= tтабл(10)=2,2281* и уровня значимости *α=0,05*.

6) Пересчитать уравнение множественной регрессии используя только значимые факторы.

7) Проверить адекватность регрессионной модели (полученной на предыдущем этапе анализа).

8) Осуществить прогнозирование в соответствии с вариантом

9) Оформить отчет о проделанной работе используя распечатки отчета, полученного средствами пакета STATISTICA или в MS Word.

**Порядок выполнения задания**

В системе STATISTICA для построения корреляционной матрицы можно воспользоваться модулем Basic Statistics/Tables (Основные статистики и таблицы), выбрав процедуры → , используя в качестве переменных все исходные данные (Select all). И процедуру для представления матрицы в графическом виде.

По корреляционной матрице можно в первом приближении судить о тесноте связи факторных признаков *х1, х2,…,xm* между собой и с результативным признаком *y*, 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.

Выбор уравнения модели, в большинстве случаев, производятся среди функций перечисленных в таблице 3. В системе STATISTICA для построения линейного уравнения множественной регрессии можно воспользоваться модулем множественной регрессии , определив зависимую (dependent) переменную *y* и независимые (independent) переменные *х1, х2, x3, x4*.

Статистический вывод о пригодности (значимости) уравнения регрессии в системе Statistica обычно проверяется в следующей последовательности.

1. Проводится общаяпроверка модели, целью которой является выяснение, объясняют ли *х*-переменные значимую долю изменения *у*. Определение значимости модели рекомендуется проводить по следующим методам (см. табл. 5).

Таблица 5

|  |  |  |
| --- | --- | --- |
| **Критерий Фишера** | **Использование уровня****значимости α** | **Использование коэффициента детерминации R2** |
| Проверяется нулевая гипотеза *H0*о равенстве полученных коэффициентов регрессии нулю: *a0=a1=a2=…=am=0*. Для этого рассчитанное системой Statistica значение *F*-критерия (*Fрасч*), сравнивается с табличным значением *Fтабл*, определяемым с использованием специальных таблиц по заданным уровню значимости (например, *α=0,05*) и числу степеней свободы (*df1=m, df2=n-m-1*). Если выполняется неравенство *Fрасч < Fтабл*, то с уверенностью, например на 95 %, можно утверждать, что рассматриваемая зависимость *y = а0 + a1x1+ … +amxm* является статистически значимой. | Если рассчитанное в Statistica значение уровня значимости *р* больше, чем заданный уровень значимости *α* (например, *α=0,05)*, то полученный результат нужно трактовать как незначимый(для 95% вероятности). В том случае, когда величина *р*<0,05, то вывод такой: это значимое уравнение с вероятностью 95%. | Рассчитанная системой Statistica величина сравнивается с табличными (критическими) значениями , определяемым с использованием специальных таблиц по заданному уровню значимости (например*, α =0,05*). Если окажется, что >, то с упомянутой степенью вероятности (95 %) можно утверждать, что анализируемая регрессия является значимой. |

Если регрессия неявляется значимой, то говорить больше не о чем.

В при веденном примере модель значима, т.к. вычисленный уровень значимости модели *р*=0,000000<0,05.

коэффициент множественной корреляции

коэффициент детерминации *R2*

Имя зависимой переменной (y) и число наблюдений

Число

степеней

свободы *df1=4,*

 *df2=10*

Вычисленный уровень значимости

Переход к результатам регрессии

Заданный уровень значимости *α*

Переход к Residual Analysis (Остаточный анализ)

Осуществив переход к результатам регрессии (Summary: Regression results) получаем уравнение линейной множественной регрессии вида *y(x1, x2, x3, x4)=6,9+0,07x1 –0,00035x2–2,08x3+0,00003x4*:

Вычисленные значения t-критерия

Вычисленные уровни

значимости

Найденные коэффициенты регрессии

коэффициент регрессии *а0* значим

коэффициент регрессии *а1*  значим

коэффициент регрессии *а3*  значим

2. Если регрессия оказывается значимой, то существует взаимосвязь между параметром *у* и переменными *х1, х2,…,xm*. Однако остается неясно, каково влияние конкретных факторов *х1, х2,…,xm* на исследуемую функцию *у*. Можно продолжить анализ, используя *t-*тесты для отдельныхкоэффициентов регрессии *а0*, *a1,a2,…,am* с целью выяснить, насколько значимой является влияние той или иной переменной *х* на параметр *у* при условии, что все другие факторы *хk* остаются неизменными. Проверку на адекватность коэффициентов регрессии рекомендуется проводить по следующим эквивалентным методам (см. табл. 5).

Таблица 5

|  |  |
| --- | --- |
| **Использование t-критерия Стьюдента** | **Использование****уровня значимости** **α** |
| Анализируемый коэффициент *а0*, *a1,a2,…,am*считается значимым, если рассчитанное системой Statistica для него значение *t*-критерия по абсолютной величине превышает *tтабл*, определяемым с использованием специальных таблиц по заданным уровню значимости (например, *α=0,05*) и числу степеней свободы (*df=n-m-1*).  | Коэффициент регрессии *а0*, *a1,a2,…,am*признается значимым, если рассчитанное системой Statistica для него значение уровня значимости *р* меньше (или равно) 0,05 (для 95%-ной доверительной вероятности). |

Т.к. вычисленные уровни значимости *p-level* для коэффициентов, стоящих при *x2* и *x4* меньше 0,05, то они не значимы. К аналогичному выводу можно прийти, воспользовавшись *t*-критерием: *t2(10)=-0,013<2,228* и *t3(10)=1,44<2,228.*

С учетом этого факта, пересчитаем уравнение множественной регрессии, выбрав в качестве зависимой (*dependent*) переменную *y* и независимые *(independent*) переменные *х1* и *x3,* коэффициенты при которых значимы:

Получаем:

Т.о., уравнение регрессии имеет вид

*y(x1, x3)=4,957+0,096x1–1,559x3*

Для выполнения прогнозов по полученному уравнению необходимо показать, что регрессионная модель адекватна результатам наблюдений. С этой целью можно воспользоваться критерием Дарбина-Уотсона, согласно которого, рассчитанный системой Statistica коэффициент *dрасч* необходимо сравнить с табличным значением *dтабл* (для совокупности объемом *n*=15, уровня значимости *α=*0,05 и трех оцениваемых параметров регрессии, значение *dтаб*л=1,75). Если *dрасч*>*dтабл*, то полученная модель адекватна и пригодна для прогнозирования. Для определения *dрасч* в Statistica в окне *Residual Analysis* на вкладке *Advanced* необходимо выбрать опцию *Durbin-Watson statistic*:

В рассматриваемом примере *dрасч=1,2<1,75*, следовательно, модель не желательно использовать для прогнозирования.

В случае, когда модель адекватна результатам наблюдения для выполнения прогноза в окне *Multyple Regression Results* вкладки *Residuals/assumptions/prediction* (*Остатки/Предположения/Прогнозирование*) выбрать опцию (прогнозирование зависимой переменной). Например, если в Москве среднегодовую стоимость основных фондов(переменная *x1*) повысить на 50 тыс. руб., а трудоемкость единицы продукции (переменная *х3*) уменьшить в два раза, то следует ожидать производительности труда равной 19,16 (увеличится на 19,16-14=5,16):

Прогнозируемое значение *y*


# 2.2 Практическое задание 2. Кластерный анализ в STATISTICA

**Постановка задачи**

Двадцать банков, акции которых котируются на рынке, предоставили следующую информацию (см. табл.), где – *x* затраты за прошлый период, *y* – прибыль за прошлый период.

Необходимо:

1) дополнить таблицу до 20 значений. Данные можно не просто придумать, а взять из любых примеров деятельности банков того или иного города, приведенных в книгах по статистике, эконометрике, СМИ, Internet или любых иных источников.

2) построить график по исходным данным (Scatterplot)

3) c использованием системы STATISTICA выяснить (дать рекомендацию) акции каких банков некоторому предприятию имеет смысл приобрести, каких – придержать, а от каких – избавиться.

Таблица

|  |  |  |
| --- | --- | --- |
| Номербанка | Затраты*x* | Прибыль*y* |
| 1 | 4 | 2 |
| 2 | 6 | 10 |
| 3 | 5 | 7 |
| 4 | 12 | 3 |
| 5 | 17 | 4 |
| 6 | 3 | 10 |
| 7 | 6 | 1 |
| 8 | 6 | 3 |
| 9 | 15 | 1 |
| 10 | 15 | 4 |
| 11 | 5 | 4 |
| 12 | 3 | 8 |
| 13 | 13 | 5 |
| 14 | 15 | 3 |
| 15 | 5 | 9 |

**Порядок выполнения задания**

**Кластерный анализ** – один из методов статистического многомерного анализа, предназначенный для группировки (кластеризации) совокупности элементов, которые характеризуются многими факторами, и получения однородных групп (кластеров). Задача кластерного анализа состоит в представлении исходной информации об элементах в сжатом виде без ее существенной потери.

STATISTICA предлагает несколько методов кластерного анализа. В дальнейшем будем использовать **Joining** (**tree clustering**) – группу иерархических методов (7 видов), которые используются в том случае, если число кластеров заранее неизвестно.

Используемый метод – **Ward’s method** – метод Уорда, который хорошо работает с небольшим количеством элементов и нацелен на выбор кластеров с примерно одинаковым количеством членов. В качестве метрики расстояния пакет предлагает различные меры, но наиболее употребительными являются **Euclidean distance** (евклидово расстояние). При кластеризации элементов в пакете STATISTICAследует выбирать режим: cases (rows) – строки, а при кластеризации факторов: variables (columns) – столбцы. В качестве переменных для рассматриваемого примере следует выбрать все переменные (all).

Для вывода результатов на экран следует выбрать

 либо .

Вывести график на печать.

Проанализировать результат и заполнить таблицу.

|  |  |  |  |
| --- | --- | --- | --- |
| Номербанка | Затраты*x* | Прибыль*y* | Рекомендацияприобрести/придержать/избавиться |
| 1 | 4 | 2 |  |
| 2 | 6 | 10 |  |
| 3 | 5 | 7 |  |
| 4 | 12 | 3 |  |
| 5 | 17 | 4 |  |
| 6 | 3 | 10 |  |
| 7 | 6 | 1 |  |
| 8 | 6 | 3 |  |
| 9 | 15 | 1 |  |
| 10 | 15 | 4 |  |
| 11 | 5 | 4 |  |
| 12 | 3 | 8 |  |
| 13 | 13 | 5 |  |
| 14 | 15 | 3 |  |
| 15 | 5 | 9 |  |
| 16 |  |  |  |
| 17 |  |  |  |
| 18 |  |  |  |
| 19 |  |  |  |
| 20 |  |  |  |