МИНИСТЕРСТВО СЕЛЬСКОГО ХОЗЯЙСТВА РОССИЙСКОЙ ФЕДЕРАЦИИ

ДЕПАРТАМЕНТ НАУЧНО-ТЕХНОЛОГИЧЕСКОЙ ПОЛИТИКИ И ОБРАЗОВАНИЯ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«КРАСНОЯРСКИЙ ГОСУДАРСТВЕННЫЙ АГРАРНЫЙ УНИВЕРСИТЕТ»

Институт подготовки кадров высшей квалификации

Паршуков Д. В.

УЧЕБНО-МЕТОДИЧЕСКИЙ КОМПЛЕКС ДИСЦИПЛИНЫ

Моделирование и статистическая обработка результатов исследования

для подготовки
аспирантов по программе
ФГОС ВПО

Курс 1, семестр 2

Квалификация выпускника - Исследователь. Преподаватель-исследователь.

Красноярск, 2015

Содержание

МОДУЛЬ 1. СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ 3
Тема 1.1 Математическая статистика 3
1.1.1 Статистическая обработка экспериментальных данных 3
1.1.2 Парная и множественная регрессия: условия и порядок построения, анализ и направления использования 14
Тема 1.2 Многомерный статистический анализ 24
1.2.1 Многомерный факторный анализ 24
1.2.2 Дискриминантные модели 37
ПРАКТИКУМ ПО МОДУЛЮ 1 49
Практическое занятие №1 Статистическая обработка данных с применением Excel 49
Практическое занятие №2 Построение и анализ моделей парной, нелинейной, множественной регрессии 57
Практическое занятие №3 Построение факторных моделей..... 64
Практическое занятие № 4 Построение дискриминантных моделей 79
МОДУЛЬ 2. ОСНОВЫ МОДЕЛИРОВАНИЯ СИСТЕМ 90
Тема 2.1 Модели прогнозирования 90
2.1.1 Классические модели прогнозирования 90
2.1.2 Нейросетевое прогнозирование..... 106
Тема 2.2 Основы имитационного моделирования 117
2.2.1 Моделирование сложных систем 117
2.2.2 Линейное и нелинейное программировани..... 134
ПРАКТИКУМ ПО МОДУЛЮ 1 144
Практическое занятие №5 Прогнозирование на основе адаптивных моделей 144
Практическое занятие № 6 Построение прогноза по нейросети 149
Практическое занятие №7 Объектно-событийное моделирование сложных систем 161
Практическое занятие № 8 Метод Монте-Карло 162
Вопросы к зачету..... 165
ЛИТЕРАТУРА 166

Модуль 1. Статистический анализ данных

Тема1.1 Математическая статистика

Лекция №1 Статистическая обработка экспериментальных данных.

План лекции

1. Генеральная и выборочная совокупности.
3. Точечные оценки неизвестных параметров.
3. Точечная оценка вероятности события

1. Генеральная и выборочная совокупности.

Для обнаружения закономерностей, описывающих исследуемое массовое явление, необходимо иметь опытные данные, полученные в результате обследования соответствующих объектов, отображающих массовое явление. Каждой генеральной совокупности соответствует случайная величина, определяемая изучаемым признаком объекта. Так как понятия генеральной совокупности и соответствующей случайной величины связаны с наблюдениями (измерениями) в неизменных условиях, то для ее обозначения (по аналогии с курсом теории вероятностей) будем использовать прописные буквы латинского алфавита (например, X, Y).

Часть отобранных объектов из генеральной совокупности называется выборочной совокупностью или выборкой.

Результаты измерений изучаемого признака n объектов выборочной совокупности порождают n значений $x_1, x_2, ..., x_n$ случайной величины X. Число n называется объемом выборки.

Наряду с генеральной совокупностью X будем рассматривать n независимых случайных величин, обозначаемых той же буквой, что и генеральная совокупность, и имеющих точно такое же распределение, как генеральная совокупность. Итак, $X_1, X_2, ..., X_n$ – n независимых экземпляров X. Если $F(x)$ – функция распределения генеральной совокупности X, то у каждой случайной величины X_i функция распределения также равна $F(x)$. Понятно, что получить n значений случайной величины X все равно что получить одно

значение n-мерной случайной величины ($X_1, X_2, ..., X_n$). Поэтому каждую выборку $x_1, x_2, ..., x_n$ объема n мы можем рассматривать как одно значение n-мерной случайной величины ($X_1, ..., X_n$).

Для того чтобы по отобранным значениям некоторого количественного показателя можно было достаточно уверенно судить обо всей совокупности, полученная выборка должна быть репрезентативной (представительной), т.е. правильно отражать пропорции генеральной совокупности. Репрезентативность выборки обеспечивается случайностью отбора объектов в выборку.

После получения (тем или иным способом) выборочной совокупности все ее объекты обследуются по отношению к определенной случайной величине – т.е. обследуемому признаку объекта. В результате этого получают наблюдаемые данные, которые представляют собой множество расположенных в беспорядке чисел. Анализ таких данных весьма затруднителен, и для изучения закономерностей полученные данные подвергаются определенной обработке.

Простейшей операцией является ранжирование опытных данных, результатом которого являются значения, расположенные в порядке неубывания. После проведения операции ранжирования опытные данные объединяют так, чтобы в каждой группе значения случайной величины были одинаковы. Значение случайной величины, соответствующее отдельной группе сгруппированного ряда наблюдаемых данных, называется вариантом, а изменение этого значения – варьированием. Варианты будем обозначать строчными буквами с соответствующими порядковому номеру группы индексами $x^{(1)}, x^{(2)}, ..., x^{(m)}$, где m – число групп. При этом имеет место $x^{(1)} < x^{(2)} < ... < x^{(m)}$.

Численность отдельной группы сгруппированного ряда данных называется частотой n_i , где i – индекс варианта, а отношение частоты данного варианта к общей сумме частот называется частностью (или относительной частотой) и обозначается ω_i , $i = 1, ..., m$, т.е.

$$\omega_i = \frac{n_i}{\sum_{i=1}^{m} n_i}. \qquad (1.1)$$

Дискретным вариационным рядом называется ранжированная совокупность вариантов $x^{(i)}$ с соответствующими им частотами n_i или частностями ω_i.

Как правило, частичные интервалы, на которые разбивается весь интервал варьирования, имеют одинаковую длину и представимы в виде

$$[z_i, z_i + h), \quad i = 1,2, ..., m, \qquad (1.2)$$

где m – число интервалов.

Длину h следует выбирать так, чтобы построенный ряд не был громоздким, но, в тоже время, позволял выявлять характерные изменения случайной величины.

Рекомендуется для h использовать следующую формулу:

$$h = \frac{x_{\max} - x_{\min}}{1 + 3.222 \lg n},$$

где $x_{\max}, x_{\min}$ – наибольшее и наименьшее значения случайной величины. Если окажется, что h – дробное число, то за длину интервала следует принять либо ближайшую простую дробь, либо ближайшую целую величину. При этом необходимо выполнение условий:

$$z_1 \le x_{\min}; \quad z_m + h \ge x_{\max}. \tag{1.3}$$

После нахождения частных интервалов определяется сколько значений случайной величины попало в каждый конкретный интервал. При этом в интервал включают значения большие или равные нижней границе и меньшие верхней границы.

3. Функция и плотность распределения. Гистограмма вариационного ряда.

В теории вероятностей для характеристики распределения случайной величины X служит функция распределения

$$F(x) = P(X < x),$$

равная вероятности события $\{X < x\}$, где x – любое действительное число.

Одной из основных характеристик выборки является выборочная (эмпирическая) функция распределения

$$F_n^*(x) = \frac{n_x}{n}, \tag{1.4}$$

где n_x – количество элементов выборки, меньших чем x.

Свойства функции $F_n^*(x)$:

1. $0 \le F_n^*(x) \le 1$; (1.5)

3. $F_n^*(x)$ – неубывающая функция;

3. $F_n^*(-\infty) = 0; \quad F_n^*(\infty) = 1.$

В качестве оценки плотности распределения вероятности непрерывной случайной величины используют гистограмму относительных частот.

Гистограммой относительных частот называется система прямоугольников, каждый из которых основанием имеет i-й интервал интервального вариационного ряда; площадь, равную относительной частоте ω_i, а высота y_i определяется по формуле

$$y_i = \frac{\omega_i}{h_i}, \quad i = 1,2,...,m,$$

где $h_i = z_{i+1} - z_i$ – длина i-го частичного интервала.

Выборочным средним $\overline{X}_в$ называется случайная величина, определенная формулой

$$\overline{X}_в = \frac{X_1 + X_2 + ... + X_n}{n}. \tag{1.6}$$

Так как конкретная выборка $x_1,...,x_n$ является реализацией значений случайных величин $X_1,...,X_n$, то среднее значение выборки

$$\bar{x}_в = \frac{x_1 + x_2 + ... + x_n}{n} \tag{1.7}$$

является одной из реализаций случайной величины $\bar{X}_в$. Другими словами, $\bar{x}_в$ есть одно из значений случайной величины $\bar{X}_в$.

Если данные представлены в виде вариационного ряда, то целесообразно для вычисления выборочного среднего одно из следующих соотношений:

- для дискретного вариационного ряда

$$\bar{x}_в = \frac{\sum_{i=1}^{m} x^{(i)} n_i}{\sum_{i=1}^{m} n_i} = \sum_{i=1}^{m} x^{(i)} \omega_i ; \tag{1.8}$$

- для интервального вариационного ряда

$$\bar{x}_в = \frac{\sum_{i=1}^{m} z_i^* n_i}{\sum_{i=1}^{m} n_i} = \sum_{i=1}^{m} \omega_i z_i^* , \tag{1.9}$$

где ω_i – частность (относительная частота), соответствующая i-й варианте или i-му частичному интервалу; z_i^* – середина i-го частичного интервала, т.е.

$$z_i^* = \frac{(z_i + z_{i+1})}{2}, \quad i = 1,2, ..., m.$$

Сравним математическое ожидание дискретной случайной величины X, вычисляемой по формуле

$$M(X) = \sum_{i=1}^{m} x_i p_i , \tag{1.10}$$

и значение выборочного среднего, определяемое (8). Прежде всего, очевидна их внешняя схожесть. Однако в формуле (10) x_i – возможные значения случайной величины, а p_i – вероятности. В формуле (8) $x^{(i)}$ – варианты случайной величины, полученные в результате наблюдений, ω_i – их относительная частота.

Так как значение выборочного среднего есть выборочный аналог математического ожидания, то имеет смысл ввести характеристику, которая бы оценивала величину рассеивания значений $x_1, x_2,..., x_n$ относительно $\bar{x}_в$, а именно

$$d_в = \sum_{i=1}^{n} \frac{(x_i - \bar{x}_в)^2}{n} . \tag{1.11}$$

Число $d_в$ является значением случайной величины

$$D_в = \sum_{i=1}^{n} \frac{(X_i - \bar{X}_в)^2}{n} , \tag{1.12}$$

которую мы будем называть выборочной дисперсией.

Если данные представлены в виде вариационного ряда, то целесообразно для вычислений $d_в$ вместо (11) использовать следующие соотношения:

- для дискретного вариационного ряда

$$d_в = \frac{\sum_{i=1}^{m}(x^{(i)} - \bar{x}_в)^2 n_i}{n} = \sum_{i=1}^{m}(x^{(i)} - \bar{x}_в)^2 \omega_i ; \qquad (1.13)$$

- для интервального вариационного ряда

$$d_в = \frac{\sum_{i=1}^{m}(z_i^* - \bar{x}_в)^2 n_i}{n} = \sum_{i=1}^{m}(z_i^* - \bar{x}_в)^2 \omega_i , \qquad (1.14)$$

где ω_i, z_i^* – те же, что и в формулах (8), (9).

3. Точечные оценки неизвестных параметров

Большинство случайных величин, рассмотренных в курсе теории вероятностей, имели распределения, зависящие от одного или нескольких параметров. Так, биномиальное распределение зависит от параметров p и n, нормальное – от параметров a и σ, распределение Пуассона – от параметра λ и т.п. Одной из основных задач математической статистики (см. разд. 1) является оценивание этих параметров по наблюдаемым данным, т.е. по выборочной совокупности.

Обозначим через θ некоторый неизвестный параметр генеральной совокупности, а через θ_n^* – точечную оценку этого параметра. Оценка θ_n^* есть функция $\varphi(X_1, X_2, ..., X_n)$ от n независимых экземпляров $X_1, X_2, ..., X_n$ генеральной совокупности, где n – объем выборки (см. пункт 3.1). Поэтому оценка θ_n^*, как функция случайных величин, также является случайной и свойства θ_n^* можно исследовать с использованием понятий теории вероятностей.

В общем случае точечная оценка θ_n^* не связана с оцениваемым параметром θ. Поэтому естественно потребовать, чтобы θ_n^* была близка к θ. Это требование формируется в терминах несмещенности, состоятельности и эффективности.

Оценка θ_n^* параметра θ называется несмещенной, если для любого фиксированного объема выборки n математическое ожидание оценки равно оцениваемому параметру, т.е.

$$M(\theta_n^*) = \theta . \qquad (1.15)$$

Оценка θ_n^* называется состоятельной, если

$$\theta_n^* \xrightarrow{p} \theta ,$$

т.е. для любого $\varepsilon > 0$ при $n \to \infty$

$$P\left(\left|\theta_n^* - \theta\right| < \varepsilon\right) \to 1 . \qquad (1.16)$$

Поясним смысл этого предельного соотношения. Пусть ε – очень малое положительное число. Тогда (3.2) означает, что чем больше число наблюдений n, тем больше уверенность (вероятность) в незначительном отклонении θ_n^* от неизвестного параметра θ. Очевидно, что «хорошая оценка» должна быть состоятельной, иначе эта оценка не имеет практического смысла, так как увеличение объема исходной информации не будет приближать нас к «истинному» значению θ.

Предположим, что генеральная совокупность распределена по нормальному закону с параметрами a и σ, причем a – математическое ожидание, подлежащее оценке, а σ^2 – известная дисперсия. Оказывается, что для любой несмещенной регулярной оценки a^* имеет место неравенство

$$D(a^*) \geq \frac{\sigma^2}{n}, \qquad (1.17)$$

где n – объем выборки, по которой производится оценивание. Если в качестве a^* принять $\overline{X}_в$, то дисперсия этой оценки, как будет показано ниже, равна $\frac{\sigma^2}{n}$, т.е. $\overline{X}_в$ – эффективная оценка параметра а, так как для нее достигается нижняя грань в неравенстве (17).

Рассмотрим на примере понятие эффективной в данном классе оценки. Предположим, что один и тот же предмет, истинная величина которого равна l, измеряется n раз различными приборами, имеющими различную точность. Пусть X_i – результаты i-го измерения. Тогда

$$M(X_i) = l, \quad D(X_i) = \sigma^2,$$

если считать, что измерения проводятся без систематических ошибок. Дисперсия σ_i^2 характеризует точность измерений. Для оценки истинного значения параметра l рассмотрим класс линейных оценок, т.е. оценок вида

$$l^* = c_1 X_1 + ... + c_n X_n,$$

где $c_1, ..., c_n$ – некоторые неизвестные константы. Из всех несмещенных оценок данного класса нужно выбрать ту, которая имеет наименьшую дисперсию.

Из несмещенности оценок получим

$$M(l^*) = M(\sum_{i=1}^{n} c_i X_i) = \sum_{i=1}^{n} c_i M(X_i) = l \sum_{i=1}^{n} c_i .$$

Значит,

$$\sum_{i=1}^{n} c_i = 1. \qquad (1.18)$$

Пользуясь свойствами дисперсии и независимостью проведенных измерений, получим

$$D(l^*) = \sum_{i=1}^{n} c_i^2 \sigma_i^2 .$$

Числа $c_1, ..., c_n$ должны удовлетворять условию (1.18) и обеспечивать минимум функции

$$F(c_1, ..., c_n) = \sum_{i=1}^{n} c_i^2 \sigma_i^2 .$$

Мы получим задачу на условный экстремум, которую можно решить с помощью функции Лагранжа:

$$L(c_1, ..., c_n) = F(c_1, ..., c_n) - \lambda(\sum_{i=1}^{n} c_i - 1).$$

Найдем критические точки функции Лагранжа:

$$\frac{\partial L}{\partial c_i} = 2c_i\sigma_i^2 - \lambda = 0, \quad i = 1, ..., n;$$

$$\sum_{i=1}^{n} c_i - 1 = 0.$$

Отсюда находим значение c_i

$$c_i = \frac{\frac{1}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}, \quad i = 1, ..., n. \tag{1.19}$$

Полученный результат имеет простой физический смысл: чем меньше точность данного прибора, тем с меньшим значением коэффициента результат его должен входить в оценку.

Математическое ожидание $M(X)$ генеральной совокупности X назовем генеральной средней $\bar{x}_г$, т.е.

$$\bar{x}_г = M(X).$$

Теорема 1. Выборочное среднее $\bar{X}_в$ есть состоятельная и несмещенная оценка генеральной средней $\bar{x}_г$.

Доказательство. Вначале покажем, что $\bar{X}_в$ есть состоятельная оценка для $\bar{x}_г$, т.е.

$$\frac{X_1 + X_2 + ... + X_n}{n} \xrightarrow{p} \bar{x}_г .$$

По следствию из теоремы Чебышева для одинаково распределенных случайных величин имеем

$$\frac{X_1 + X_2 + ... + X_n}{n} \xrightarrow{p} M(X).$$

Так как $M(X) = \bar{x}_г$, то используя свойства математического ожидания, получим

$$M(\bar{X}_в) = M\left(\frac{X_1 + ... + X_n}{n}\right) = \frac{M(X_1) + ... + M(X_n)}{n} =$$

$$= \frac{nM(X)}{n} = \bar{x}_г .$$

Теорема доказана.

Теорема 3. Пусть случайная величина X имеет нормальное распределение $N(a,\sigma)$, где a – математическое ожидание, σ^2 – дисперсия

случайной величины X. Тогда выборочное среднее $\overline{X}_в$ является эффективной несмещенной оценкой для $\bar{x}_г$.

Доказательство. Необходимо показать, что дисперсия $D(\overline{X}_в)$ совпадает с минимальной дисперсией, равной в случае нормального распределения σ^2/n, а ее математическое ожидание $M(\overline{X}_в)$ равно $\bar{x}_г$.

Найдем дисперсию $D(\overline{X}_в)$:

$$D(\overline{X}_в) = D(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{nD(X)}{n^2} = \frac{\sigma^2}{n}. \qquad (1.20)$$

Мы проверили при доказательстве теоремы 1, что $M(\overline{X}_в) = \bar{x}_г$. Так как дисперсия $D(\overline{X}_в)$ равна минимальному значению, то выборочное среднее $\overline{X}_в$ является эффективной несмещенной оценкой. Теорема доказана.

Таким образом показано, что выборочное среднее $\overline{X}_в$ имеет все три свойства «хорошей» оценки и этим объясняется ее широкое использование в качестве оценки математического ожидания генеральной совокупности.

Напомним, что по конкретной выборке $x_1, ..., x_n$ вычисляется (см. (1.7), (1.8), (1.9)) «конкретное значение» $\bar{x}_в$, являющееся одним из множества возможных значений случайной величины $\overline{X}_в$.

Дисперсию $D(X)$ генеральной совокупности X будем называть генеральной дисперсией $D_г$, т.е.

$$D_г = D(X). \qquad (1.21)$$

Теорема 3. Выборочная дисперсия $D_в$ является состоятельной, но смещенной оценкой генеральной дисперсии $D_г$.

Доказательство. Получим сначала формулу для вычисления $D_в$. Согласно определению

$$D_в = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_в)^2}{n}.$$

С другой стороны,

$$\sum_{i=1}^{n}(X_i - \overline{X}_в)^2 = \sum_{i=1}^{n}(X_i^2 - 2\overline{X}_в X_i + \overline{X}_в^2) =$$

$$= \sum_{i=1}^{n} X_i^2 - 2n\overline{X}_в^2 + n\overline{X}_в^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}_в^2.$$

Тогда из определения дисперсии следует

$$D_в = \frac{\sum_{i=1}^{n} X_i^2 - n\overline{X}_в^2}{n} = \frac{\sum_{i=1}^{n} X_i^2}{n} - \overline{X}_в^2.$$

Воспользуясь теперь следствием из теоремы Чебышева для одинаково распределенных случайных величин X_i^2 и свойствами предела по вероятности, получаем

$$\frac{\sum_{i=1}^{n} X_i^2}{n} \xrightarrow{p} M(X_i^2) = M(X^2);$$

$$\overline{X}_в \xrightarrow{p} M(X)$$

и, значит,

$$D_в \xrightarrow{p} M(X^2) - M^2(X) = D(X) = D_г.$$

Следовательно, дисперсия выборочная $D_в$ является состоятельной оценкой для дисперсии генеральной. Вычислим математическое ожидание $D_в$ и убедимся, что $M(D_в) \neq D_г$. Имеем

$$M(D_в) = M\left(\frac{\sum_{i=1}^{n} X_i^2}{n} - \overline{X}_в^2\right) = M\left(\frac{\sum_{i=1}^{n} X_i^2}{n}\right) - M(\overline{X}_в^2) =$$

$$= M\left(\frac{\sum_{i=1}^{n} X_i^2}{n}\right) - M\left(\frac{X_1 + \ldots + X_n}{n}\right)^2 =$$

$$= M(X^2) - M\left(\frac{X_1^2 + X_2^2 + \ldots + X_n^2 + \sum_{i \neq j} X_i X_j}{n^2}\right),$$

где $\sum_{i \neq j} X_i X_j$ означает сумму произведений величин X_i и X_j для всех значений i и j от 1 до n, но не равных между собой. Так как X_i и X_j независимы при $i \neq j$, то

$$M(X_i X_j) = M(X_i) M(X_j).$$

Поэтому, продолжая вычисления $M(D_B)$, получаем

$$M(D_в) = M(X^2) - \frac{M(X_1^2 + \ldots + M(X_n^2) + \sum_{i \neq j} M(X_i) M(X_j)}{n^2} =$$

$$= M(X^2) - \frac{nM(X^2) + n(n-1)M^2(X)}{n} =$$

$$= M(X^2) - \frac{nM(X^2) + n(n-1)}{n} = \frac{n-1}{n}\left[M(X^2) - M^2(X)\right]$$

$$= \frac{n-1}{n} D_г.$$

Множитель $n(n-1)$ объясняется тем, что по правилу произведения количество различных пар (i, j) при $1 \leq i \neq j \leq n$ равно $n(n-1)$.

Итак, мы получили, что

$$M(D_в) = \frac{n-1}{n} D_г \qquad 1.22)$$

и, следовательно, $D_в$ – смещенная оценка для дисперсии генеральной. Теорема доказана.

Полученная формула (22) для вычисления математического ожидания дисперсии выборочной позволяет указать состоятельную и несмещенную оценку для дисперсии генеральной. Для этого рассмотрим случайную величину

$$S^2 = \frac{n}{n-1} D_в, \qquad (1.23)$$

называемую исправленной дисперсией. Понятно, что

$$S^2 \xrightarrow{p} D_г,$$

так как $\frac{n}{n-1} \to 1$ при $n \to \infty$. С другой стороны,

$$M(S^2) = M\left(\frac{n}{n-1} D_в\right) = \frac{n}{n-1} M(D_в) = \frac{n}{n-1} \cdot \frac{n-1}{n} D_г = D_г.$$

Тем самым доказана.

Теорема 4. Исправленная дисперсия S^2 является состоятельной и несмещенной оценкой для дисперсии генеральной $D_г$.

Заметим, что для выборок большого объема множитель $\frac{n}{n-1}$ близок к 1 и поэтому случайные величины S^2 и $D_в$ мало отличаются друг от друга. Однако для выборок малого объема это отличие может быть существенным.

Возникает вопрос: будет ли несмещенная оценка S^2 эффективной.

Для ответа предположим, что случайная величина X подчиняется нормальному распределению $N(a,\sigma)$, а величины $X_1, X_2, ..., X_n$, как обычно, n независимых экземпляров независимой величины X. Тогда минимальная дисперсия несмещенной оценки для дисперсий равна

$$D_{\min} = \frac{2\sigma^4}{n}. \qquad (24)$$

Величина S^2 представима в виде

$$S^2 = \frac{\sigma^2}{n-1} \chi^2_{n-1}, \qquad (1.24)$$

где χ^2_{n-1} – случайная величина, имеющая χ^2 – распределение с $n-1$ степенями свободы. Поэтому

$$D(S^2) = \frac{\sigma^4}{(n-1)^2} D(\chi^2_{n-1}) = \frac{2\sigma^4}{n-1}, \qquad (1.25)$$

из этого следует

$$D(S^2) = \frac{n}{n-1} D_{\min}. \qquad (1.26)$$

Следовательно, S^2, будучи несмещенной оценкой дисперсии $D(X)$, не является эффективной оценкой. Однако при достаточно больших n увеличение $D(S^2)$ по сравнению с $D_{\min}$ пренебрежительно мало.

Заметим, что несмещенная эффективная оценка дисперсии $D(X)$ нормально распределенной величины $X = N(a,\sigma)$ имеет вид

$$S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2.$$

Однако в эту формулу входит математическое ожидание a, которое, как правило, заранее неизвестно.

3. Точечная оценка вероятности события

Обозначим через $p(A)$ неизвестную вероятность события A в одном испытании. Для оценивания $p(A)$ проведем n независимых испытаний, в которых событие A произошло m раз. Тогда случайная величина

$$p^* = \frac{m}{n} \tag{1.27}$$

является частностью (относительной частотой) события A. Свойства этой точечной оценки определяет.

Теорема 3. Относительная частота $p^* = m/n$ появления события A в n испытаниях есть состоятельная оценка вероятности $p(A)$.

Доказательство. Состоятельность оценки p^* вытекает из теоремы Бернулли, согласно которой для любого $\varepsilon > 0$ выполняется неравенство

$$\lim_{n\to\infty} P\left(\left|\frac{m}{n} - P(A)\right| < \varepsilon\right) = 1 \tag{1.28}$$

или в других обозначениях:

$$\frac{m}{n} \xrightarrow{p} p(A).$$

Для доказательства несмещенности этой оценки зафиксируем число испытаний n. Найдем математическое ожидание частности m/n, имея в виду, что в условиях испытаний Бернулли величина т имеет биномиальный закон распределения с характеристиками М(т) = пр, D(m) = пр(1 – р). Имеем

$$M\left(\frac{m}{n}\right) = \frac{1}{n}M(m) = \frac{1}{n}np = p(A).$$

Следовательно, $p^* = m/n$ является несмещенной оценкой вероятности р(А).

Для доказательства эффективности укажем, что минимум среди дисперсий различных несмещенных оценок вероятности р(А) равен

$$D_{\min} = \frac{p(1-p)}{n}. \tag{1.29}$$

Определим дисперсию оценки p^*

$$D(p^*) = D\left(\frac{m}{n}\right) = \frac{1}{n^2}D(m) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

Так как D(P*) СОВпадает с минимальной дисперсией $D_{\min}$, то частность р*, будучи несмещенной оценкой, является также и эффективной.

Вопросы по лекции 1

1. Что называют статистической/генеральной совокупностью?
3. Понятие и виды выборки.
3. Понятие вариационного ряда?
4. Описание алгоритма построения непрерывного вариационного ряда.
5. Понятие и свойства эмпирической функцией распределения.
6. Формулы нахождения выборочной средний статистического распределения.
7. Определение выборочной дисперсии и ее назначение.
8. Записать формулы для вычисления дисперсии для простой и взвешенной выборки.
9. Записать формулы для вычисления исправленной дисперсии и рассказать для чего она вводится.
10. Понятие моды и медианы вариационного ряда.
11. Алгоритм вычисления математического ожидания и дисперсии по методу произведений.
13. Дать определения асимметрии и эксцесса статистического распределения и рассказать об их назначении.

Лекция №2 Парная и множественная регрессия: условия и порядок построения, анализ и направления использования.

План лекции

1. Понятие регрессионного анализа
3. Парная линейная регрессия
3. Проверка качества уравнения регрессии
4. Множественная регрессия.

1. Понятие регрессионного анализа

В естественных науках большей частью имеют дело со строгими (функциональными) зависимостями, при которых каждому значению одной переменной соответствует единственное значение другой. Однако в подавляющем большинстве случаев между экономическими переменными таких зависимостей нет. Поэтому в экономике говорят не о функциональных, а о корреляционных либо статистических зависимостях.

Если переменные обозначить X и Y, то зависимость вида:

$$M(Y|X) = f(x) \qquad (3.1)$$

называется функцией регрессии Y на X. При этом X называется независимой (объясняющей) переменной (прегрессором), Y – зависимой (объясняемой) переменной. При рассмотрении двух случайных величин говорят о парной регрессии.

Зависимость нескольких переменных, выражаемую функцией

$$M(Y|\ x_1, x_2, \ldots, x_m) = f(x_1, x_2, \ldots, x_m), \qquad (3.2)$$

называют множественной регрессией.

Под регрессией понимается функциональная зависимость между объясняющими переменными и условным математическим ожиданием (средним значением) зависимой переменной, которая строится с целью предсказания (прогнозирования) этого среднего значения при фиксированных значениях первых.

Для отражения того факта, что реальные значения зависимой переменной не всегда совпадают с ее условными математическими ожиданиями и могут быть различными при одном и том же значении объясняющей переменной (наборе объясняющих переменных), фактическая зависимость должна быть дополнена некоторым слагаемым ε, которое, по существу, является случайной величиной и указывает на стохастическую суть зависимости. Из этого следует, что связи между зависимой и объясняющей (ими) переменными выражаются соотношениями

$$Y = M(Y|X) + \varepsilon; \qquad (3.3)$$

$$Y = M(Y|\ x_1, x_2, \ldots, x_m) + \varepsilon, \qquad (3.4)$$

называемыми регрессионными моделями (уравнениями).

Выбор формулы связи переменных называется спецификацией уравнения регрессии. В случае парной регрессии выбор формулы обычно осуществляется по графическому изображению реальных статистических данных в виде точек в декартовой системе координат, которое называется корреляционным полем (диаграммой рассеивания) (рис. 1.1).

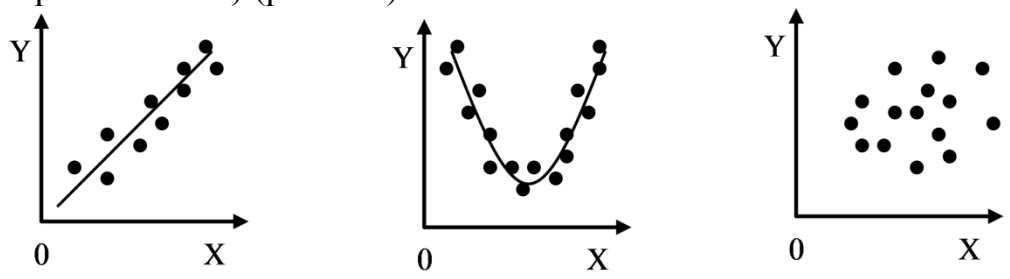


Рис. 1

Если на первых двух графиках относительно четко определяется форма связи (для первого – линейная, для второго – квадратичная), то для третьего явная взаимосвязь между переменными отсутствует. В случае множественной регрессии определение подходящего вида регрессии является наиболее сложной задачей.

3. Парная линейная регрессия

Если функция регрессии линейна, то говорят о линейной регрессии. Модель линейной регрессии (линейное уравнение) является наиболее распространенным (и простым) видом зависимости между экономическими переменными. Кроме того, построенное линейное уравнение может служить начальной точкой эконометрического анализа. Линейная регрессия (теоретическое линейное уравнение регрессии) представляет собой линейную функцию между условным математическим ожиданием $M(Y|X=x_i)$ зависимой переменной Y и одной объясняющей переменной X (x_i – значения независимой переменной в i-ом наблюдении, $i=1,2,\ldots,n$).

$$M(Y|X=x_i)=\beta_0+\beta_1 x_i\,. \qquad (3.5)$$

Для отражения того факта, что каждое индивидуальное значение y_i отклоняется от соответствующего условного математического ожидания, необходимо ввести в последнее соотношение случайное слагаемое ε_i.

$$y_i=M(Y|X=x_i)+\varepsilon_i=\beta_0+\beta_1 x_i+\varepsilon_i. \qquad (3.6)$$

Это соотношение называется теоретической линейной регрессионной моделью, β_0 и β_1 – теоретическими параметрами (теоретическими коэффициентами) регрессии, ε_i – случайным отклонением.

Следовательно, индивидуальные значения y_i представляются в виде суммы двух компонент – систематической $(\beta_0+\beta_1 x_i)$ и случайной ε_i, причина появления которой достаточно подробно рассмотрена ранее. В общем виде теоретическую линейную регрессионную модель будем представлять в виде:

$$Y=\beta_0+\beta_1 X+\varepsilon\,. \qquad (3.7)$$

Для определения значений теоретических коэффициентов регрессии необходимо знать и использовать все значения переменных X и Y генеральной совокупности, что практически невозможно.

Таким образом, задачи линейного регрессионного анализа состоят в том, чтобы по имеющимся статистическим данным (x_i, y_i), $i = 1,2,\ldots,n$ для переменных X и Y:

а) получить наилучшие оценки неизвестных параметров β_0 и β_1;

б) проверить статистические гипотезы о параметрах модели;

в) проверить, достаточно ли хорошо модель согласуется со статистическими данными (адекватность модели данным наблюдений).

Следовательно, по выборке ограниченного объема мы сможем построить так называемое эмпирическое уравнение регрессии

$$\hat{y}_i = b_0 + b_1 x_i, \qquad (3.8)$$

где $\hat{y}_i$ – оценка условного математического ожидания $M(Y|X = x_i)$; b_0 и b_1 – оценки неизвестных параметров β_0 и β_1, называемые эмпирическими коэффициентами регрессии. Следовательно, в конкретном случае:

$$y_i = b_0 + b_1 x_i + e_i, \qquad (3.9)$$

где отклонение e_i – оценка теоретического случайного отклонения ε_i.

Для оценки неизвестных параметров можно использовать следующие подходы:

1. $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)$, однако эта сумма не может быть мерой качества найденных оценок в силу того, что существует бесчисленное количество прямых, для которых $\sum_{i=1}^{n} e_i = 0$.

2. $\sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} \lfloor y_i - \hat{y}_i \rfloor = \sum_{i=1}^{n} |y_i - b_0 - b_1 x_i|$. Этот метод называется методом наименьшей суммы.

3. $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$. Это самый распростаренный и теоретически обоснованный метод, который получил название метода наименьших квадратов (МНК). Кроме того, он является наиболее простым с вычислительной точки зрения.

Параметры уравнения регрессии могут быть найдены из следующих выражений:

$$\begin{cases} b_1 = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2}; \\ b_0 = \bar{y} - b_1 \bar{x}. \end{cases} \qquad (3.9)$$

Доказано, что для получения по МНК наилучших результатов необходимо, чтобы выполнялся ряд предпосылок относительно случайного отклонения (предпосылки Гаусса-Маркова):

A. Математическое ожидание случайного отклонения ε_i равно нулю: $M(\varepsilon_i)=0$ для всех наблюдений.

B. Дисперсия случайных отклонений ε_i постоянна: $D(\varepsilon_i)=D(\varepsilon_j)=\sigma^2$ для любых наблюдений i и j. Данное условие подразумевает, что несмотря на то, что при каждом конкретном наблюдении случайное отклонение может быть либо большим, либо меньшим, не должно быть некой причины, вызывающей большую ошибку (отклонение).

C. Случайные отклонения ε_i и ε_j являются независимыми друг от друга для $i \neq j$. Выполнимость данной предпосылки предполагает, что отсутствует систематическая связь между любыми случайными отклонениями.

D. Случайное отклонение должно быть независимо от объясняющих переменных.

E. Модель является линейной относительно параметров.

3. Проверка качества уравнения регрессии

В силу того, что случайные отклонения ε_i по выборке определены быть не могут, при анализе надежности оценок коэффициентов регрессии они заменяются отклонениями $e_i = y_i - b_0 - b_1 x_i$ значений y_i переменной Y от оцененной линии регрессии. Дисперсия случайных отклонений $D(\varepsilon_i)=\sigma^2$ заменяется ее несмещенной оценкой

$$S^2 = \frac{1}{n-2}\sum (y_i - b_0 - b_1 x_i)^2 = \frac{\sum e_i^2}{n-2}. \qquad (3.10)$$

Тогда

$$D(b_0) \approx S_{b0}^2 = \frac{S^2 \sum x_i^2}{n\sum (x_i - \bar{x})^2} = \overline{x^2} S_{b1}^2. \qquad (3.11)$$

$$D(b_1) \approx S_{b1}^2 = \frac{S^2}{\sum (x_i - \bar{x})^2}. \qquad (3.12)$$

$S^2 = \frac{\sum e_i^2}{n-2}$ – необъясненная дисперсия (мера разброса зависимой переменной вокруг линии регрессии).

$S = \sqrt{\frac{\sum e_i^2}{n-2}}$ – стандартная ошибка оценки (стандартная ошибка регрессии).

$S_{b0} = \sqrt{S_{b0}^2}$ и $S_{b1} = \sqrt{S_{b1}^2}$ – стандартные отклонения случайных величин b_0 и b_1, называемые стандартными ошибками коэффициентов регрессии.

Проверка гипотез относительно коэффициентов линейного уравнения регрессии

При проведении статистического анализа перед исследователем зачастую возникает необходимость сравнения эмпирических коэффициентов регрессии b_0 и b_1 с некоторыми теоретически ожидаемыми значениями β_0 и β_1 этих коэффициентов. Данный анализ осуществляется при помощи статистической проверкой гипотез.

Для проверки гипотезы $b_1 = \beta_1$ используется статистика:

$$t = \frac{b_1 - \beta_1}{S_{b1}}, \tag{3.13}$$

которая при справедливости $b_1 = \beta_1$ имеет распределение Стьюдента с числом степеней свободы $v = n - 2$, где n – объем выборки. Следовательно, гипотеза $b_1 = \beta_1$ отклоняется на основании данного критерия, если:

$$|T_{набл}| = \left| \frac{b_1 - \beta_1}{S_{b1}} \right| \geq t_{\frac{\alpha}{2},n-2}, \tag{3.14}$$

где α – требуемый уровень значимости. При невыполнении (3.15) считается, что нет оснований для отклонения $b_1 = \beta_1$.

Наиболее важной на начальном этапе статистического анализа построенной модели является задача установления значимости коэффициентов. Эта задача решается при помощи отношений, называемых t-статистикой:

$$t = \frac{b_1}{S_{b1}} \qquad \text{и} \qquad t = \frac{b_0}{S_{b0}}. \tag{3.16}$$

В случае, если $|t| > t_{\frac{\alpha}{2},n-2}$, то статистическая значимость соответствующего коэффициента регрессии подтверждается. Значения $t_{\frac{\alpha}{2},n-2}$ в зависимости от уровня значимости и числа степеней свободы v ($v = n - 2$) приведены в приложении 1.

При оценке значимости коэффициента линейной регрессии на начальном этапе можно использовать следующее «грубое» правило (почти всегда работает при $n > 10$), позволяющее не прибегать к таблицам:

- Если $|t| \leq 1$, то коэффициент не может быть признан значимым, так как доверительная вероятность составит менее, чем 0,7.
- Если $1 < |t| \leq 2$, то найденная оценка может рассматриваться как относительно (слабо) значимая. Доверительная вероятность в этом случае лежит между 0,7 и 0,93.
- Если $2 < |t| \leq 3$, то доверительная вероятность составляет от 0,95 до 0,99.
- Если $|t| > 3$, то это почти гарантия значимости коэффициента.

Интервальные оценки коэффициентов линейного уравнения регрессии

Доверительные интервалы коэффициентов b_0 и b_1, которые с надежностью $(1-\alpha)$ накрывают определяемые параметры β_0 и β_1, определяются по формулам:

$$\left(b_0 - t_{\frac{\alpha}{2},n-2} S(b_0);\ b_0 + t_{\frac{\alpha}{2},n-2} S(b_0) \right) \quad \text{и} \quad \left(b_1 - t_{\frac{\alpha}{2},n-2} S(b_1);\ b_1 + t_{\frac{\alpha}{2},n-2} S(b_1) \right). \tag{3.17}$$

Доверительные интервалы для зависимой переменной

Интервал, определяющий границы, за пределами которых могут оказаться не более $100\alpha\%$ точек наблюдений при $X = x_p$, рассчитывается следующим образом:

$$\left(b_0 + b_1 x_p \pm t_{\frac{\alpha}{2},n-2} S \sqrt{1 + \frac{1}{n} + \frac{(\bar{x} - x_p)^2}{\sum (x_i - \bar{x})^2}} \right). \qquad (3.18)$$

4. Множественная регрессия.

Если на исследуемый объект показатель оказывает влияние не один, а несколько факторов, то вместо парной регрессии $M(Y|X) = f(x)$ рассматривается множественная регрессия

$$M(Y|x_1, x_2, \ldots, x_m) = f(x_1, x_2, \ldots, x_m). \qquad (3.19)$$

Задача оценки статистической взаимосвязи переменных Y и $X_1, X_2, \ldots, X_m$ формулируется аналогично случаю парной регрессии. Уравнение множественной регрессии может быть представлено в виде:

$$Y = f(\beta, X) + \varepsilon, \qquad (3.20)$$

где $X = (X_1, X_2, \ldots, X_m)$ – вектор независимых (объясняющих) переменных; β – вектор параметров (подлежащих определению); ε – случайная ошибка (отклонение); Y – зависимая (объясняемая) переменная.

Рассмотрим самую употребляемую и наиболее простую из моделей множественной регрессии – модель множественной линейной регрессии.

Теоретическое линейное уравнение регрессии имеет вид:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m + \varepsilon \qquad (3.21)$$

или для индивидуальных наблюдений $i, i = 1,2,\ldots,n$:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_m x_{im} + \varepsilon_i. \qquad (3.22)$$

Здесь $\beta = (\beta_0, \beta_1, \ldots, \beta_m)$ – вектор размерности $(m+1)$ неизвестных параметров. $\beta_j, j = 1,2,\ldots,m$ называется j -тым теоретическим коэффициентом регрессии (частичным коэффициентом регрессии). Он характеризует чувствительность величины Y к изменению величины X_j, т.е. отражает влияние на условное математическое ожидание $M(Y|x_1, x_2, \ldots, x_m)$ зависимой переменной Y объясняющей переменной X_j при условии, что все другие объясняющие переменные модели остаются постоянными. β_0 – свободный член, определяющий Y в случае, когда все объясняющие переменные X_j равны нулю.

Данные наблюдений и соответствующие коэффициенты в матричной форме выглядят следующим образом:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1m} \\ 1 & x_{21} & x_{22} & \ldots & x_{2m} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nm} \end{pmatrix}, \quad B = \begin{pmatrix} b_0 \\ b_1 \\ \ldots \\ b_m \end{pmatrix}, \quad e = \begin{pmatrix} e_1 \\ e_2 \\ \ldots \\ e_n \end{pmatrix}.$$

Здесь $Y-n$-мерный вектор-столбец наблюдений зависимой переменной X; X – матрица размерности $n\times(m+1)$, в которой i-тая строка $(i=1,2,\ldots,n)$ представляет наблюдение вектора значений независимых переменных $X_1, X_2, \ldots, X_m$; единица соответствует переменной при свободном члене b_0; B – вектор-столбец размерности $(m+1)$ параметров уравнения регрессии; e – вектор-столбец размерности n отклонений выборочных (реальных) значений y_i зависимой переменной Y от значений $\hat{y}_i$, получаемых по уравнению регрессии

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_m x_{im}. \qquad (3.23)$$

Функция $Q=\sum_{i=1}^{n} e_i^2$ в матричной форме представима как произведение вектор-строки $e^T=(e_1,e_2,\ldots,e_n)$ на вектор-столбец e. Вектор-столбец e может быть в свою очередь представлен в следующем виде:

$$e = Y - XB. \qquad (3.24)$$

Отсюда:

$$Q = e^T e = (Y-XB)^T (Y-XB) = Y^T Y - B^T X^T Y - Y^T XB + B^T X^T XB =$$
$$= Y^T Y - 2B^T X^T Y + B^T X^T XB.$$

Здесь e^T, B^T, X^T, Y^T – векторы и матрицы, транспонированные к e, B, X, Y соответственно. При выводе формулы использовались следующие известные соотношения линейной алгебры:

$$(Y-XB)^T = Y^T - (XB)^T; \quad (XB)^T = B^T X^T; \quad B^T X^T Y = Y^T XB.$$

Необходимым условием экстремума функции Q является равенство нулю ее частных производных $\frac{\partial Q}{\partial b_j}$ по всем параметрам $b_j, j=1,2,\ldots,m$. Вектор-столбец $\frac{\partial Q}{\partial B}$ частных производных в матричном виде выглядит следующим образом:

$$\frac{\partial Q}{\partial B} = -2X^T Y + 2(X^T X)B. \qquad (3.25)$$

Рассмотрим более подробно нахождение $\frac{\partial Q}{\partial B}$. Очевидно, что

$$\frac{\partial Q}{\partial B} = \frac{\partial (Y^T Y)}{\partial B} - 2\frac{\partial (B^T X^T Y)}{\partial B} + \frac{\partial (B^T X^T XB)}{\partial B}.$$

$Y^T Y$ от B не зависит, следовательно, $\frac{\partial (Y^T Y)}{\partial B} = 0$.

Обозначим вектор-столбец $X^T Y$ размерности $(m+1)$ через R. Тогда $B^T X^T Y = B^T R = \sum_{j=0}^{m} b_j r_{j+1}$, где r_{j+1} – соответствующий элемент вектора R. Поэтому $\frac{\partial (B^T R)}{\partial B} = R = X^T Y$.

Обозначим матрицу $X^T X$ размерности $(m+1)\times(m+1)$ через Z. Тогда

$$S = B^T ZB = \left((b_0, b_1, \ldots, b_m) \cdot \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1,m+1} \\ z_{21} & z_{22} & \cdots & z_{2,m+1} \\ \cdots & \cdots & \cdots & \cdots \\ z_{m+1,1} & z_{m+1,2} & \cdots & z_{m+1,m+1} \end{pmatrix} \right) \cdot \begin{pmatrix} b_0 \\ b_1 \\ \cdots \\ b_m \end{pmatrix} =$$

$$= \left(\sum_{i=0}^{m} b_i z_{i+1,1}, \sum_{i=0}^{m} b_i z_{i+1,2}, \ldots, \sum_{i=0}^{m} b_i z_{i+1,m+1} \right) \cdot \begin{pmatrix} b_0 \\ b_1 \\ \cdots \\ b_m \end{pmatrix} = \sum_{j=0}^{m} b_j \sum_{i=0}^{m} b_i z_{i+1,j+1} = = \sum_{j=0}^{m} \sum_{i=0}^{m} b_j b_i z_{i+1,j+1}.$$

Следовательно, частная производная $\frac{\partial S}{\partial b_j} = 2\sum_{i=0}^{m} b_i z_{i+1,j+1}$.

В результате имеем $\frac{\partial S}{\partial B} = 2(X^T X)B$.

Следовательно, формула (3.12) справедлива. Приравняв $\frac{\partial Q}{\partial B}$ к нулю, получаем:

$$-2X^T Y + 2(X^T X)B = 0 \Rightarrow$$
$$X^T Y = (X^T X)B \Rightarrow$$
$$B = (X^T X)^{-1} X^T Y.$$

Здесь $(X^T X)^{-1}$ – матрица, обратная к $X^T X$.

Построение эмпирического уравнения регрессии является начальным этапом эконометрического анализа. Первое же построенное по выборке уравнение регрессии очень редко является удовлетворительным по тем или иным характеристикам. Поэтому следующей важнейшей оценкой является проверка качества уравнения регрессии. В эконометрике принята устоявшаяся схема такой проверки, которая проводится по следующим направлениям:

- проверка статистической значимости коэффициентов уравнения регрессии;
- проверка общего качества уравнения регрессии;
- проверка свойств данных, выполнимость которых предполагалась при оценивании уравнения (проверка выполнимости предпосылок МНК).

Вопросы по лекции 3.

1. Понятие корреляционной зависимости между двумя признаками X и Y.
3. Определение и формула нахождения условной средней признака Y.
3. Область и условия применения теории корреляции.
4. Формулы нахождения параметров уравнения парной регрессии.
5. Формула и интерпретация коэффициента корреляции.
6. Дать определение коэффициента линейной корреляции, сформулировать его свойства.
7. Оценка значимости уравнения регрессии и ее параметров.
8. Как определяется значимость коэффициента линейной корреляции?
9. Записать доверительные интервалы для оценки коэффициента линейной корреляции при различных объемах выборки.

10. Записать формулу для нахождения коэффициента детерминации в случае парной линейной корреляции и рассказать о его назначении.

11. Рассказать о проверке адекватности уравнения линейной регрессии y на x для случая несгруппированных опытных данных.

13. Ошибка аппроксимации и формула для ее нахождения.

Тема 1.2 Многомерный статистический анализ

Лекция №3 Многомерный факторный анализ

План лекции

1. Общий алгоритм факторного анализа
3. Метод главных компонент
3. Применение МГК для решения задач классификации.

1. Общий алгоритм факторного анализа

В большом количестве случаев, исследователь решая задачу описания исследуемого объекта и создания статистических моделей сталкивается с трудностями при определении числа оценочных характеристик, показателей, критериев и т.д. При описании сложных объектов большое число показателей может при статистической обработке привести к некорректным результатам. Для снижения исходной размерности рассматриваемых признаков, а также более качественного и детального описания модели исследуемого явления или объекта применяют факторный анализ.

Определим следующую задачу:

Исходными данными является таблица наблюдений Y, в столбцах которой представлены объекты, измеренные в количественной шкале (j = 1,2,…,n), а строки содержат описания объектов по некоторым критериям (i = 1,2,…,N). элементы x_{ij} на пересечении строк и столбцов есть количественное значение i-го критерия для j-го объекта.

Необходимо выделить новые независимые критерии для оценки, количеством l (l<N), из множества первоначальных критериев путем разделения первоначального множества на некоторые подмножества. Причем данные подмножества должны быть независимы между собой, а внутри подмножеств, связаны общими свойствами.

В теории измерений известно, что измеренная величина содержит по крайней мере, две компоненты: истинное значение переменной и ошибку измерения

$$x_{ij} = x_{il} + x_{ik} + x_{is}$$

В качестве способа выделения данных подмножеств будем использовать факторный анализ. Под факторным анализом понимается совокупность статистических моделей, описывающих и объясняющих наблюдаемые данные с помощью небольшого числа скрытых (латентных) факторов, которые могут быть сконструированы с помощью определенных математических методов. Модели факторного анализа применяются при решении следующих задач:

- редукция данных или понижение размерности признакового пространства типа "объект – признак" за счет сведения многочисленных взаимозависимых наблюдаемых переменных к некоторым обобщенным ненаблюдаемым факторам;

- преобразование исходных переменных к виду, более удобному для визуализации или интерпретации; классификация объектов на основе сжатого признакового пространства;

- создание структурной теории исследования объектов и интерпретация косвенных факторов, не поддающихся непосредственному измерению.

Факторный анализ основан на предположении, что связи исходных признаков – это результат воздействия сравнительно небольшого числа неявных, т.е. латентных, факторов (конструктов). Так как число латентных факторов обычно значительно меньше числа исходных признаков, то основной задачей факторного анализа считается сжатие исходного массива переменных. Факторный анализ основывается на матрице корреляций между переменными. Приведем описание факторного анализа.

Под моделью факторного анализа понимают представление исходных переменных в виде линейной комбинации факторов F

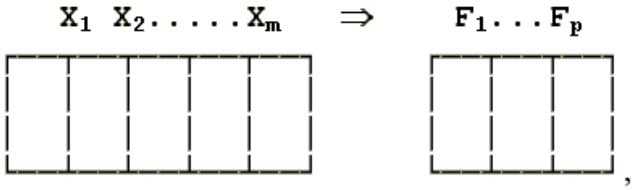


Рис. 3.1 Факторный анализ.

рассчитанных так, чтобы наилучшим способом (с минимальной погрешностью) представить X:

$$X_i = \sum_{k=1}^{l} a_{ik} F_k + U_i \qquad (3.1)$$

В этой модели латентные переменные F_k, k = 1,2,…,l, называются общими факторами, а переменные U_i , i = 1,2,…,N, – специфическими факторами (англ. – unique factor). Значения a_{ik} называются факторными нагрузками.

Основное требование к исходным данным для факторного анализа – это то, что они должны подчиняться допущению о многомерном нормальном распределении в совокупности. Для проверки этой гипотезы используют тест "сферичности" распределения данных Бартлетта, где оценивается предположение о диагональности матрицы корреляций. Если эта гипотеза не отвергается (т.е. наблюдаемый уровень значимости превышает 5%) – нет смысла в факторном анализе, поскольку направления главных осей случайны. На практике предположение о многомерной нормальности проверить весьма трудно, поэтому факторный анализ чаще всего применяется без такой процедуры, тем более, что ряд исследователей считает это допущение излишним.[] В нашем случае на входе мы подбираем выборку таким образом, чтобы она соответствовала закону нормального распределения путем проверки с помощью теста Колмогорова - Смирнова.

Если предполагается, что все признаки X_j стандартизованы $\sigma_i = 1$, $m(X_j) = 0$), а факторы F_1, F_2, …, F_p независимы и не связаны со специфическими факторами U_j, то факторные нагрузки a_{jk} совпадают с коэффициентами корреляции между общими факторами и переменными X_j. Общая дисперсия признака X_j раскладывается при этом на сумму квадратов факторных нагрузок H_i^2, которая называется общностью, и дисперсию специфического фактора $S_{u_i}^2$, или специфичность:

$$S_{x_i}^2 = H_i^2 + S_{U_i}^2 \text{ , где } \quad H_i^2 = \sum_k a_{ik} \tag{3.2}$$

Другими словами, общность H_i^2 представляет собой часть дисперсии переменных, объясненную факторами, а специфичность $S_{u_i}^2$ – часть дисперсии, обусловленную случайными ошибками или переменными, неучтенными в модели. В соответствии с постановкой задачи, необходимо искать такие факторы, при которых суммарная общность максимальна, а специфичность – минимальна.

Основным объектом преобразований является корреляционная матрица из коэффициентов корреляции Пирсона (иногда – это может быть дисперсионно-ковариационная матрица), полученная обычным путем обработки массива данных Y. Выделение общих факторов и сжатие информации сводится к воспроизведению с той или иной степенью точности исходной корреляционной матрицы, т.е. предполагается, что редуцированная корреляционная матрица получена с использованием тех же объектов, но описанных меньшим числом переменных. Фактически под сжатием информации понимается уменьшение размерности корреляционной матрицы, а не самих данных, тем более что восстановить исходные данные по корреляционной матрице нельзя.

Коэффициенты, составляющие корреляционную матрицу, могут вычисляться разным способом. Различают следующие техники факторного анализа:

R –техника, когда коэффициенты корреляции вычисляются между переменными и исходная матрица X сжимается по строкам, т.е. число признаков уменьшается с N до l;

Q –техника, когда изучается корреляция между объектами (точнее, их состояниями, описываемыми векторами параметров) и их количество уменьшается с n до p;

P – техника, предполагающая факторный анализ результатов экспериментальных исследований, выполненных на одном и том же объекте в различные промежутки времени.

Для построения модели риска банкротства предлагается применять R – технику.

Для поиска факторов будем использовать один из методов факторного анализа − метод главных компонент. Его основное различие от факторного анализа заключается в том, что главные компоненты F_k связаны с наблюдаемыми переменными X_j линейными функциями преобразования:

$$X_j = \sum_{k=1}^{p} a_{jk} F_k \tag{3.3}$$

$$F_k = \sum_{j=1}^{m} a_{jk} X_j \tag{3.4}$$

Метод главных компонент более прост в расчетах и интерпретации, но одна из главных трудностей его использования – необходимость преобразования исходных данных, представленных в разных единицах измерения, в сопоставимые величины. Традиционным методом преобразования является нормирование по стандартным отклонениям, когда матрица Z стандартизованных исходных данных определяется по формуле $z_{ij} = \frac{x_{ij} - x_j}{S_j}$, где x_j – среднее значение j-го признака; S_j – стандартное отклонение; j = 1,2,…,m; i = 1, 2,…,n.

Для вычисления корреляционной матрицы R размером m´ m, имеет место простое матричное соотношение:

$$R = \frac{1}{m-1} ZZ^T \tag{3.5}$$

где T – символ транспонирования.

Основная идея метода главных компонент основана на следующем предположении: чем выше дисперсия вдоль какой-нибудь оси, тем больше информации содержат значения проекций на эту ось. Поэтому вполне естественно предпринять попытку отыскать ось с максимальной дисперсией, которую можно было бы рассматривать как "ординационную". Такая ось называется первой главной компонентой (фактором).

Поиск всей системы взаимно перпендикулярных осей по методу главных компонент будет сводится к последовательной процедуре: ищем первый фактор, который объясняет наибольшую часть дисперсии, ищем независимый от первого фактора, второй фактор, объясняющий наибольшую часть оставшейся дисперсии, и т.д.

Геометрически это выглядит следующим образом (см. рис. 3.2).

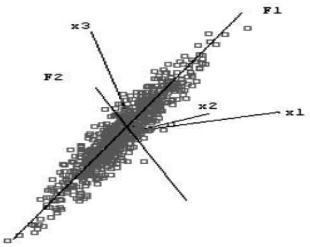


Рис. 3.2. Сжатие признакового пространства с применением факторного анализа

Для построения первого фактора F_1 берется прямая, проходящая через центр координат и облако рассеяния данных. При этом отыскивается такая ось, для которой сумма квадратов расстояний всех точек до перпендикуляра к этой прямой была бы максимальна. Это означает, что этой осью объясняется максимум дисперсии переменных. Найденная ось после нормировки используется в качестве первого фактора. Если "облако" данных вытянуто в виде эллипсоида (имеет форму "огурца"), фактор F_1 совпадет с направлением, в котором вытянуты объекты, и по нему с наибольшей точностью можно предсказать значения исходных переменных. Для поиска второго фактора F_2 ищется ось, перпендикулярная первому фактору, также объясняющая наибольшую часть дисперсии, не объясненной первой осью. После нормировки эта ось становится вторым фактором. Если данные представляют собой плоский элипсоид ("блин") в трехмерном пространстве, два первых фактора позволяют в точности описать эти данные. Максимально возможное число главных компонент равно количеству переменных.

Основная модель метода главных компонент записывается в матричном виде следующим образом:

$$Z = A*F, \qquad (3.6)$$

где Z – матрица m´ n стандартизованных исходных данных; A – матрица m´ p факторных нагрузок (факторное отображение); F – матрица p´ n значений факторов; m – количество переменных и n – количество объектов исходной матрицы; p – количество выделенных факторов. Очевидно, что неизвестными являются матрицы A и F.

Вычислительные аспекты метода главных компонент сводятся к следующим шагам:

Решается характеристическое матричное уравнение

$$R = L *V \qquad (3.7)$$

которое в общем случае имеет m корней l , называемых собственными или характеристическими числами корреляционной матрицы R, каждому из которых соответствует вектор-столбец V базисных функций. Собственными значениями квадратной матрицы R порядка m называются такие значения 1_k, при которых система следующих m уравнений имеет нетривиальное решение:

$$RV_k = 1_k V_k, \qquad (3.8)$$

где V_k – собственные векторы матрицы R, соответствующие 1_k.; k = 1,2,…,m.

Из последовательности собственных значений 1_k выбирается p максимальных. Матрица факторных нагрузок A каждой исходной переменной j на каждый выделенный фактор k, соответствующая коэффициентам линейных преобразований a_{jk}, вычисляется по формуле

$$a_{jk} = v_{ik} \times (1_k)^{0,5},\ j = 1, 2, ..., p \qquad (3.9)$$

Редуцированную матрицу факторов F, соответствующую исходной таблице X наблюдений, в которой количество столбцов уменьшено с m до p, рассчитывают по формуле:

$$f_{ik} = \sum_{j=1}^{m} a_{jk} z_{ij} \qquad (3.10)$$

Основная проблема расчетов состоит в оценке того, сколько главных компонент необходимо построить для оптимального представления анализируемых исходных факторов. Величина 1_k представляет не что иное, как часть суммарной дисперсии совокупности преобразованных данных, объясненную главной компонентой F_k. Если переменные стандартизованы, то $1_{1>l_2>l_3,...}$ и необходимо иметь в виду, что первые несколько членов разложения дают основной вклад в объяснение вариации величин в исходных данных.

Решение о том, когда следует остановить процедуру выделения компонент, зависит главным образом от точки зрения на то, что считать малой долей дисперсии. Это решение достаточно произвольно, однако имеются два критерия: критерий Кайзера (Kaiser) и критерий "каменистой осыпи" Кэттелла (Cattell), которые в большинстве случаев позволяют рационально выбрать число компонент. Но анализ составляющих с малыми величинами собственных значений едва ли целесообразен еще и потому, что они могут оказаться статистически недостоверными из-за ошибок различного происхождения. Ввиду того, что иллюстративной целью факторного анализа часто является получение факторного отображения в графическом виде, обычно ограничивают p=2 и дают изображение пространства в двумерном срезе, поскольку выполнить это для трех и более выделенных факторов проблематично.

Для интерпретации факторов необходимо приписать каждому из них некоторый содержательный смысл, связанный с предметной областью. Чтобы понять, какая реальность скрыта в найденных факторах, необходимо провести анализ корреляций факторных нагрузок с исходными переменными. Для повышения интерпретируемости факторов использует метод варимаксного вращения VARIMAX, позволяющий добиться большей "выразительности" матрицы факторных нагрузок. Его суть состоит в изменении координатных осей, образуемых факторами, с целью получить более контрастные нагрузки, так называемой простой факторной структуры. Новые факторы в результате вращения осей ищутся в виде специального вида линейной комбинации имеющихся факторов: $\hat{F}_i = \sum_{k=1}^{m} b_{ik} F_k$, максимизирующей "дисперсии" квадратов факторных нагрузок для переменных

$$\sum_i \left[\sum_k a_{ik}^4 / m - \left[\sum_k a_{ik}^2 / m \right]^2 \right] \to \max \qquad (3.11)$$

Чем сильнее разойдутся квадраты факторных нагрузок к концам отрезка [0,1], тем больше будет значение целевой функции вращения и тем четче интерпретация факторов.

Составим математическую модель факторного анализа, далее именуемой FA-моделью (Model of factor analysis) по алгоритму факторного анализ:

$$FA =< F, K, A, R > \quad (3.12)$$

Здесь F- факторы, полученные в результате факторного анализа, К-набор признаков, определяющих каждый из факторов F, A - факторные нагрузки между К и F, R (restriction -ограничение) - ограничения по критериям Кайзера и Кеттлера для определения конечного числа факторов. Скобки определяют совокупность или набор значений.

2. Метод главных компонент

Пусть задана $(p \times n)$ -матрица наблюдений случайной векторной переменной $\mathbf{X} = [X_1 \ldots X_p]'$ с вектором средних $\boldsymbol{\mu}_x = [\mu_1, \ldots, \mu_p]'$ и ковариационной матрицей $К_х$, определяющей структуру зависимости между переменными $X_j,\ j = 1, \ldots, p$. Нужно найти линейное преобразование, которое позволило бы получить сжатое представление исходных данных меньшим числом переменных без существенной потери информации, содержащейся в исходной матрице. Преобразуем эти наблюдения (р×р) -ортогональной матрицей вида

$$\boldsymbol{\Phi} = [\varphi_1 \ldots \varphi_p]', \quad (3.13)$$

где $\varphi_j = [\varphi_{1j} \ldots \varphi_{pj}]', (j = 1, \ldots, p)$ – система р-мерных ортонормированных векторов, т. е. для скалярного произведения (*,*) справедливо

$$(\varphi_i, \varphi_j) = \begin{cases} 1 & \text{при} \quad i = j, \\ 0 & \text{при} \quad i \neq j. \end{cases} \quad (3.14)$$

Тогда получаем случайную векторную переменную Y с некоррелированными компонентами

$$\mathbf{Y} = [Y_1 \ldots Y_p]' = \boldsymbol{\Phi}\mathbf{X}, \quad (3.15)$$

где Y_j есть линейная комбинация координат признаков $X_j,\ j = 1, \ldots, p$

$$Y_j = \varphi_{1j} x_{j1} + \ldots + \varphi_{pj} x_{jp},\ j = 1, \ldots, p. \quad (3.16)$$

Из (2) следует, что $\boldsymbol{\Phi}\boldsymbol{\Phi}' = \boldsymbol{\Phi}'\boldsymbol{\Phi} = \mathbf{I}$ и $\boldsymbol{\Phi}' = \boldsymbol{\Phi}^{-1}$, поэтому $\mathbf{X} = \boldsymbol{\Phi}'\mathbf{Y}$ или

$$\mathbf{X} = \varphi_1 Y_1 + \ldots + \varphi_p Y_p. \quad (3.17)$$

Ковариационная матрица данных X (по определению) равна

$$\mathbf{K_x} = M\{(\mathbf{X} - \boldsymbol{\mu}_x)(\mathbf{X} - \boldsymbol{\mu}_x)'\}. \quad (3.18)$$

Определитель $|\mathbf{K_x}|$ ковариационной матрицы $\mathbf{K_x}$ называют обобщенной дисперсией матрицы данных X.

Ковариационная матрица $\mathbf{K_y}$ случайной векторной переменной Y определяется выражением

$$\begin{aligned} \mathbf{K_y} &= M\{(\mathbf{Y} - \boldsymbol{\mu}_y)(\mathbf{Y} - \boldsymbol{\mu}_y)'\} = M\{\boldsymbol{\Phi}(\mathbf{X} - \boldsymbol{\mu}_x)(\mathbf{X} - \boldsymbol{\mu}_x)'\boldsymbol{\Phi}'\} = \\ &= \boldsymbol{\Phi} M\{(\mathbf{X} - \boldsymbol{\mu}_x)(\mathbf{X} - \boldsymbol{\mu}_x)'\}\boldsymbol{\Phi}' = \boldsymbol{\Phi}\mathbf{K_x}\boldsymbol{\Phi}'. \end{aligned} \quad (3.19)$$

Так как $\mathbf{K}_x$ и $\boldsymbol{\Phi}$ являются квадратными матрицами, то определитель ковариационной матрицы $\mathbf{K_y}$ равен

$$|\mathbf{K}_y| = |\mathbf{\Phi K_x \Phi'}| = |\mathbf{\Phi\Phi'}||\mathbf{K}_x| = |\mathbf{K}_x|, \quad (3.20)$$

т. е. обобщенные дисперсии матриц X и Y равны.

Наилучшее ортогональное преобразование должно обеспечить наименьшую избыточность. Это означает, что матрица Y должна иметь некоррелированные компоненты Y_j, $j = 1,\ldots,p$. Другими словами, матрица $\mathbf{K}_y$ должна быть диагональной

$$\mathbf{K}_y = diag[\sigma^2_{y_1},\ldots,\sigma^2_{y_p}], \quad (3.12)$$

где $\sigma^2_{y_j}$ – дисперсия j-ой компоненты случайной векторной переменной Y.

Обозначим $\lambda_j = \sigma^2_{y_j}$, $j = 1,\ldots,p$. Тогда

$$|\mathbf{K}_y| = \prod_{j=1}^{p} \lambda_j. \quad (3.22)$$

Положим, что дисперсии упорядочены $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p \geq 0$. Если не все λ_j равны между собой, то матрицу Y можно сжать отбрасыванием компонент с пренебрежимо малыми дисперсиями. Пусть $Y_1 - (n \times 1)$ -вектор является первой главной компонентой матрицы X $Y_1 = \sum_{i=1}^{p} \varphi_{i1} x_{1i}$.

Найдем дисперсию этой главной компоненты

$$\sigma^2_{y_1} = \boldsymbol{\varphi}_1' \mathbf{K}_x \boldsymbol{\varphi}_1 = \sum_{r=1}^{p}\sum_{i=1}^{p} \varphi_{1i}\varphi_{r1} M[(X_1 - \mu_1)(X_1 - \mu_1)].$$

Потребуем, чтобы первая компонента Y_1 имела наибольшую дисперсию при условии сохранения ортогональности векторов $\boldsymbol{\varphi}_i$ матрицы Ф. Тогда задача нахождения наилучшего преобразования $\boldsymbol{\varphi}_1$ сводится к нахождению максимума функции $\boldsymbol{\varphi}_1' \mathbf{K}_x \boldsymbol{\varphi}_1$ при условии

$$(\boldsymbol{\varphi}_1', \boldsymbol{\varphi}_1) = \sum_{j=1}^{p} \varphi_{1j}^2 = 1.$$

Чтобы решить эту задачу оптимизации обычно вводят функцию Лагранжа

$$L(\boldsymbol{\varphi}) = \boldsymbol{\varphi}'_1 \mathbf{K}_x \boldsymbol{\varphi}_1 - \lambda_1(\boldsymbol{\varphi}'_1 \boldsymbol{\varphi}_1 - 1), \quad (3.23)$$

где λ_1 – множитель Лагранжа. Необходимое условие экстремума получим, приравняв нулю частные производные $\partial L / \partial \varphi_1$:

$$\partial L / \partial \boldsymbol{\varphi}_1 = 2(\mathbf{K}_x \boldsymbol{\varphi}_1 - \lambda_1 \boldsymbol{\varphi}_1) = 2(\mathbf{K}_x - \lambda_1 \mathbf{I})\boldsymbol{\varphi}_1 = 0, \quad (3.24)$$

где I – единичная матрица. Поскольку нас интересуют только решения, при которых $\boldsymbol{\varphi}_1 \neq 0$, то должно удовлетворяться условие на определитель

$$|\mathbf{K}_x - \lambda_1 \mathbf{I}| = 0. \quad (3.25)$$

Отсюда следует, что λ_1 есть собственное число матрицы $\mathbf{K}_x$, а $\boldsymbol{\varphi}_1$ – соответствующий этому числу собственный вектор. Выражение (3.24) может быть переписано в виде $\mathbf{K}_x \boldsymbol{\varphi}_1 = \lambda_1 \boldsymbol{\varphi}_1$.

$$\boldsymbol{\varphi}'_1 \mathbf{K}_x \boldsymbol{\varphi}_1 = \lambda_1 \boldsymbol{\varphi}'_1 \boldsymbol{\varphi}_1 = \lambda_1. \quad (3.26)$$

Левая часть равенства (3.26) есть $\sigma^2_{y_1}$, а поскольку решалась задача максимизации $\sigma^2_{y_1}$, следовательно, λ_1 есть максимальное собственное число

матрицы $\mathbf{K}_x$. Чтобы найти вторую главную компоненту $\mathbf{Y}_2 = \boldsymbol{\varphi}'_2\mathbf{X}$ потребуем выполнения двух условий – условия нормировки: $(\boldsymbol{\varphi}'_2,\boldsymbol{\varphi}_2) = \sum_{i=1}^{p}\varphi_{2i}^2 = 1$ и условия ортогональности: $(\boldsymbol{\varphi}'_2,\boldsymbol{\varphi}_2) = 0$. Вектор $\boldsymbol{\varphi}_2$ определяется теперь так, чтобы $\boldsymbol{\sigma}_{y_2}^2$ была максимальна при выполнении двух указанных условий. Эта задача требует использования двух множителей Лагранжа λ_2 и β. Мы должны максимизировать выражение

$$\boldsymbol{\varphi}'_2\mathbf{K}_x\boldsymbol{\varphi}_2 - \lambda_2(\boldsymbol{\varphi}'_2\boldsymbol{\varphi}_2 - 1) - \beta(\boldsymbol{\varphi}'_1\boldsymbol{\varphi}_2 - 1). \qquad (3.27)$$

Взяв производную от выражения (3.27) и приравняв ее к 0, находим в соответствии с условием (2), что $\beta = 0$. Учитывая условия нормировки, получаем, что λ_2 есть второе по величине собственное число матрицы $\mathbf{K}_x$, равное дисперсии второй главной компоненты $\lambda_2 = \sigma_{y_2}^2$, а $\boldsymbol{\varphi}_2$ – соответствующий собственный вектор. Процесс повторяется до тех пор, пока не будут найдены все собственные числа и ассоциированные с ними собственные векторы, которые являются дисперсиями и коэффициентами линейных комбинаций главных компонент.

Таким образом мы нашли преобразование, задаваемое ортогональной матрицей Ф, столбцы которой являются собственными векторами ковариационной матрицы $\mathbf{K}_x$.

С точки зрения геометрической интерпретации ортогональное преобразование есть вращение системы координат p-мерного векторного пространства вокруг начала координат. Суммарная дисперсия компонент векторной величины Y равна

$$\sum_{j=1}^{p}\sigma_{y_j}^2 = trM\{(\mathbf{Y}-\boldsymbol{\mu}_y)(\mathbf{Y}-\boldsymbol{\mu}_y)'\} = trM\{\boldsymbol{\Phi}(\mathbf{X}-\boldsymbol{\mu}_x)(\mathbf{X}-\boldsymbol{\mu}_x)'\boldsymbol{\Phi}'\}.$$

Используя свойство следа произведения матриц, имеем

$$\sum_{j=1}^{p}\sigma_{y_j}^2 = trM\{(\mathbf{X}-\boldsymbol{\mu}_x)(\mathbf{X}-\boldsymbol{\mu}_x)'\boldsymbol{\Phi}'\boldsymbol{\Phi})\} = tr\mathbf{K}_x = \sum_{j=1}^{p}\sigma_{x_j}^2, \text{ или}$$

$$\sum_{j=1}^{p}\lambda_j = tr\mathbf{K}_y = tr\mathbf{K}_x, \qquad (3.28)$$

где $tr\mathbf{K}_y$, $tr\mathbf{K}_x$ – следы матриц $\mathbf{K}_y$ и $\mathbf{K}_x$.

Относительный вклад компоненты Y_j в общую дисперсию случайной векторной переменной Y равен

$$\boldsymbol{\sigma}_{y_j}^2 / \sum_{j=1}^{p}\boldsymbol{\sigma}_{y_j}^2 = \lambda_j / \sum_{j=1}^{p}\lambda_j = \lambda_j / tr\mathbf{K}_x. \qquad (3.29)$$

Полученное преобразование максимизирует дисперсию первых компонент Y_j, называемых главными компонентами, что обеспечивает наилучшее сжатие. Это преобразование иногда называют преобразованием Карунева-Лоэва.

Если в преобразованном пространстве ограничиться r первыми компонентами матрицы Y, а остальные $p-r$ положить равными нулю, то получим сжатую матрицу случайной векторной переменной ${}_r\mathbf{Y}_n$ с r компонентами. Действительно, из r $(r < p)$ собственных векторов формируется $(r \times p)$ -матрица ортогонального преобразования ${}_r\boldsymbol{\Phi}_p = [\varphi_1 \ldots \varphi_r]'$, определяемая

соотношением $\mathbf{Y} = \mathbf{\Phi X}$. При этом обобщенная дисперсия уменьшится на величину $\prod_{j=r+1}^{p} \lambda_j$.

Для i-й главной компоненты имеем $Y_i = \varphi_i' X_i$, где φ_i – собственный вектор, соответствующий собственному числу λ_i матрицы $\mathbf{K}_x$. Важность i-й главной компоненты определяется ее вкладом в общую дисперсию

$$\sigma_{y_i}^2 / \sum_{j=1}^{p} \sigma_{y_j}^2 = \lambda_i / \sum_{j=1}^{p} \lambda_j. \qquad (3.30)$$

Ортогональное преобразование не изменяет общей дисперсии. Если ограничиться r первыми компонентами, то их доля в общей дисперсии составит $\sum_{i=1}^{r} \lambda_i / \sum_{j=1}^{p} \lambda_j$. Остаточная дисперсия будет равна $\sum_{j=r+1}^{p} \lambda_j$. Таким образом, дисперсия остатков равна сумме дисперсий, соответствующих отброшенным компонентам вектора Y, и это справедливо для любого ортонормированного преобразования. Приведенный критерий сжатия называется дисперсионным. Для использования этого метода нужно найти собственные числа матрицы $\mathbf{K}_x$, упорядочить их в порядке убывания, выбрать такое количество компонент r, которое обеспечит заданную долю дисперсии остатков

$$\sigma_{\delta_x}^2 = \sum_{j=r+1}^{p} \lambda_j / tr\mathbf{K}_x = \sum_{j=r+1}^{p} \lambda_j / \sum_{j=1}^{p} \lambda_j. \qquad (3.31)$$

Решение о том, когда следует остановить процедуру выделения компонент, главным образом зависит от точки зрения на то, что считать малой долей дисперсии. Это решение достаточно произвольно, однако имеются два критерия: критерий Кайзера (Kaiser) и критерий каменистой осыпи Кэттелла (Cattell), которые в большинстве случаев позволяют рационально выбрать число компонент.

При использовании критерия Кайзера отбираются только компоненты, с собственными значениями, большими единицы. По существу, это означает, что если компонента не выделяет дисперсию, эквивалентную, по крайней мере, дисперсии одной переменной, то он опускается. Этот критерий удобен при нахождении главных компонент по корреляционной матрице исходных данных и является, вероятно, наиболее широко используемым. Критерий каменистой осыпи является графическим методом. В этом случае рассматривается изображение собственных чисел, которые наносятся на график в порядке их убывания. Выделение компонент заканчивается на той компоненте, после которой исследуемая зависимость близка к горизонтальной и похожа на "каменистую осыпь" обломков горных пород, скапливающихся в нижней части скалистого склона.

Критерий Кайзера иногда сохраняет слишком много факторов, в то время как критерий каменистой осыпи иногда сохраняет слишком мало факторов; однако оба критерия дают хорошие результаты, когда имеется относительно небольшое число компонент и много переменных. На практике принимается тот критерий, для которого полученное число компонент может быть содержательно интерпретировано. Поэтому обычно исследуется несколько решений с большим или меньшим числом компонент, и затем выбирается одно наиболее "осмысленное".

3. Применение МГК для решения задач классификации.

Рассмотрим M классов $\omega_1,\ldots,\omega_M$, образы которых представлены векторами $\mathbf{X}_i$, $i=1,\ldots,M$. Каждый образ описывается p признаками. Тогда наблюдения можно представить матрицей

$$\mathbf{X}=\{x_{ij}\},\ i=1,\ldots,M,\ j=1,\ldots,p\cdot$$

Будем полагать, что известна априорная вероятность $p(\omega_i)$ появления i-го класса. При таких условиях порядок выделения признаков будет следующим:

1.По образам, входящим в обучающую выборку находим ковариационную матрицу $\mathbf{K}_x$:

$$\mathbf{K}_x=\sum_{i=1}^{M}p(\omega_i)M[\mathbf{X}_i\mathbf{X}_i'],$$

либо вычисляем ее оценку.

2.По $\mathbf{K}_x$ находим собственные числа и ортонормированные собственные векторы.

3.Из r собственных векторов формируем $(r\times p)$-матрицу ортогонального преобразования

$$\mathbf{\Phi}=[\varphi_1\ \ \ldots\ \ \varphi_r]'\cdot$$

4. Находим главные компоненты $Y_i=\mathbf{\Phi}X_i,\ i=1,\ldots,r$. Поскольку Ф – матрица размера $(r\times p)$ и X_i– p-мерный вектор, то очевидно, что компоненты Y_i при $r<p$ представляют собой образы, имеющие размерность, меньшую p.

5. Производим классификацию одним из известных методов кластерного или дискриминантного анализов.

Необходимые условия применения МГК. Для того, чтобы применение МГК приводило к получению оптимальных результатов, необходимо выполнение условия $MY_i=0$, что равносильно условию $MX_i=0$. Последнее выполняется автоматически, если отдельные классы характеризуются нулевыми математическими ожиданиями. На первый взгляд может показаться, что проблему можно решить, центрируя образы отдельных классов относительно соответствующих математических ожиданий. Однако при решении распознавания образов отсутствуют сведения относительно принадлежности образа определенному классу. Естественно, затруднения не возникают, если математические ожидания всех классов равны.

Пример 3. Этот пример заимствован из книги [5]. Пусть даны два класса образов ω_1 *и* ω_2, каждый из которых описывается тремя параметрами. Образы представлены на рис. 1 и в координатах – на странице 13.

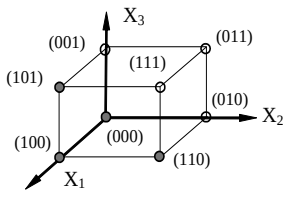


Рис. 1. Исходные образы

$$\omega_1 \qquad\qquad \omega_2$$

$$\mathbf{x}_{11}=\begin{bmatrix}0\\0\\0\end{bmatrix}, \qquad \mathbf{x}_{21}=\begin{bmatrix}0\\0\\1\end{bmatrix},$$

$$\mathbf{x}_{12}=\begin{bmatrix}1\\0\\0\end{bmatrix}, \qquad \mathbf{x}_{22}=\begin{bmatrix}0\\1\\0\end{bmatrix},$$

$$\mathbf{x}_{13}=\begin{bmatrix}1\\0\\1\end{bmatrix}, \qquad \mathbf{x}_{23}=\begin{bmatrix}0\\1\\1\end{bmatrix},$$

$$\mathbf{x}_{14}=\begin{bmatrix}1\\1\\0\end{bmatrix}, \qquad \mathbf{x}_{24}=\begin{bmatrix}1\\1\\1\end{bmatrix}.$$

Вычислим собственные числа и соответствующие им собственные векторы ковариационной матрицы $\mathbf{K}_x$

$$\lambda_1=1,\ \boldsymbol{\varphi}_1=\frac{1}{\sqrt{3}}\begin{bmatrix}1\\1\\1\end{bmatrix},\ \lambda_2=\frac{1}{4},\ \boldsymbol{\varphi}_2=\frac{1}{\sqrt{6}}\begin{bmatrix}-2\\1\\1\end{bmatrix},\ \lambda_3=\frac{1}{4},\ \boldsymbol{\varphi}_3=\frac{1}{\sqrt{2}}\begin{bmatrix}0\\1\\-1\end{bmatrix}.$$

А. Выберем собственные числа λ_1 и λ_2 и соответствующие им собственные векторы. Получим матрицу преобразования

$$\boldsymbol{\Phi}=\begin{bmatrix}\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}}\\ -\frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}}\end{bmatrix}.$$

Находим векторы главных компонент для обоих классов

$$y_{1j}=\boldsymbol{\Phi}x_{ij},\ y_{2j}=\boldsymbol{\Phi}x_{2j}$$

$$\omega_1 \qquad\qquad \omega_2$$

$$\mathbf{y}_{11}=\begin{bmatrix}0\\0\end{bmatrix}, \qquad \mathbf{y}_{21}=\frac{1}{\sqrt{6}}\begin{bmatrix}\sqrt{2}\\1\end{bmatrix},$$

$$\mathbf{y}_{12}=-\frac{1}{\sqrt{6}}\begin{bmatrix}\sqrt{2}\\-2\end{bmatrix}, \qquad \mathbf{y}_{22}=\frac{1}{\sqrt{6}}\begin{bmatrix}\sqrt{2}\\1\end{bmatrix},$$

$$\mathbf{y}_{13}=\frac{1}{\sqrt{6}}\begin{bmatrix}2\sqrt{2}\\-1\end{bmatrix}, \qquad \mathbf{y}_{23}=\frac{1}{\sqrt{6}}\begin{bmatrix}2\sqrt{2}\\2\end{bmatrix},$$

$$\mathbf{y}_{14}=\frac{1}{\sqrt{6}}\begin{bmatrix}2\sqrt{2}\\-1\end{bmatrix}. \qquad \mathbf{y}_{24}=\frac{1}{\sqrt{6}}\begin{bmatrix}3\sqrt{2}\\0\end{bmatrix}.$$

Соответствующие образы представлены на рис. 3. Разделимость объектов в сжатом пространстве сохранилась.

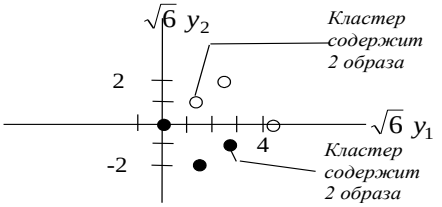


Рис. 2. Образы, сжатые до двумерного варианта

В. Рассмотрим матрицу преобразования $\mathbf{\Phi} = (1/\sqrt{3})[1\ \ 1\ \ 1]$, содержащую только один собственный вектор, соответствующий первому собственному числу. Это преобразование переводит образы в точки:

ω_1	ω_2
$y_{11} = 0,$	$y_{21} = \sqrt{3},$
$y_{12} = \sqrt{3},$	$y_{22} = \sqrt{3},$
$y_{13} = 2\sqrt{3},$	$y_{23} = 2\sqrt{3},$
$y_{14} = 2\sqrt{3}.$	$y_{24} = 3\sqrt{3}.$

Соответствующие образы представлены на рис. 3,

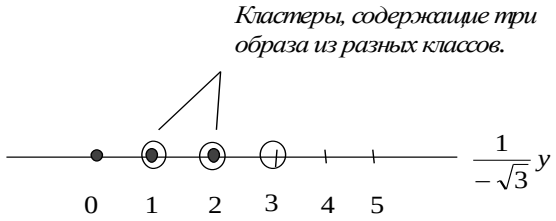


Рис. 3. Образы, сжатые до одномерного варианта

из которого видно, что образы, принадлежащие к разным классам, перекрываются, поэтому такое сжатие нежелательно.

Вопросы по лекции 3

1. Определение и цели факторного анализа.
2. Суть задачи снижения размерности данных.
3. Область применения факторного анализа.
4. Понятие и назначение матрицы факторных нагрузок.
5. Сущность метода главных компонент.
6. Сущность центроидного вектора.
7. Применение факторного анализа для моделирования временных рядов.
8. Интерпретация общностей.
9. Решение модели факторного анализа.
10. Применение прикладных программных продуктов для проведения факторного анализа.

Лекция № 4 Дискриминантные модели

План лекции

1. Понятие и область применения дискриминантного анализа.
2. Линейная дискриминантная функция.
3. Алгоритм дискриминантного анализа.

1. Понятие и область применения дискриминантного анализа.

Дискриминантный анализ (discriminant analysis) - один из методов распознавания образов (pattern recognition). Задачей распознавания образов является построение правила, позволяющего с наименьшей ошибкой предсказывать принадлежность объекта к одному из заданных классов по его описанию. Для построения этого правила используется предварительная информация о наборе объектов, для которых известны как описание, так и принадлежность к классу. Например, имеется группа пациентов, для которых известен набор симптомов и точный диагноз. Требуется, используя эти данные, построить правило, дающее возможность в дальнейшем ставить диагноз по симптомам новым пациентам.

Рассмотрим задачу распознавания образов, пользуясь вероятностными терминами. Пусть рассматриваемые потенциальные объекты (наблюдения) характеризуются значениями m переменных X_1, X_2…., X_m, являющихся непрерывными случайными величинами, распределенными по одному из законов $F_1(x_1, x_2…., x_m)$, $F_2(x_1, x_2…., x_m)$, …или $F_k(x_1, x_2…., x_m)$ в зависимости от того, к какому из классов k=1, 2, …, K относится объект.

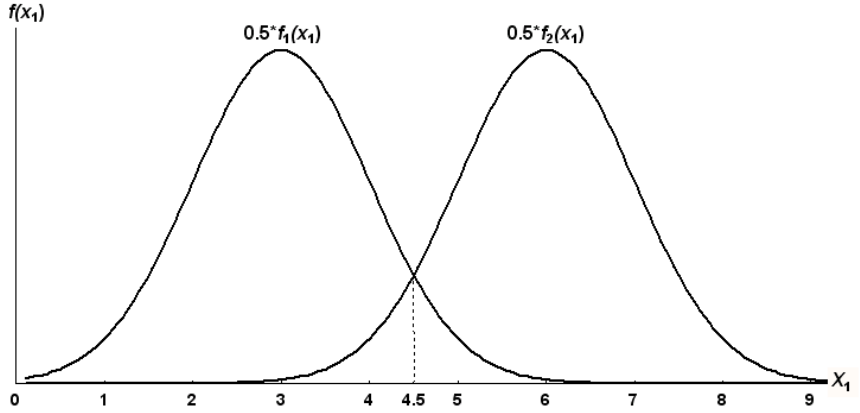


Рис. 4.1 Распознавание между двумя классами, заданными двумя одномерными нормальными распределениями с равными априорными вероятностями.

Пусть, например, имеется два класса, заданными двумя одномерными нормальными распределениями $N(\mu=3, \sigma^2=1)$ и $N(\mu=5, \sigma^2=1)$., причем объекты из этих классов встречаются с равными априорными вероятностями p=q=0.5. На рис. 4.1 изображены плотности этих двух распределений с равными весами. Мы видим, что левее точки $X_1=4.5$ плотность $f_1(x_1)$ выше плотности $f_2(x_1)$, а справа – наоборот. Правило, состоящее в том, чтобы любое новое наблюдение,

оказавшееся левее точки X_1=4.5 относить к первому классу, а оказавшееся правее – ко второму, обеспечивает минимальную ошибку классификации.

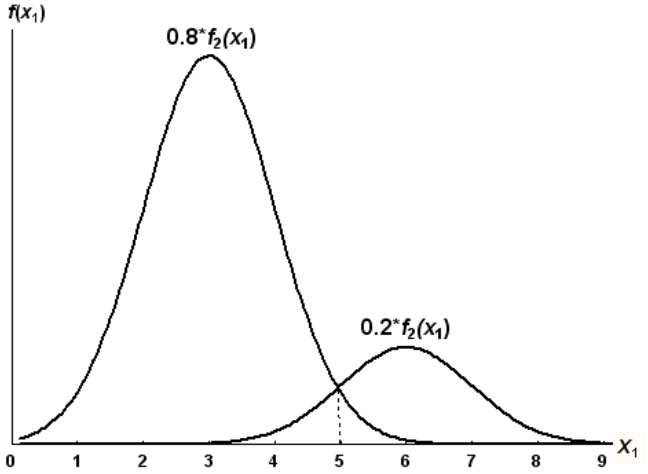


Рис. 4.2. Распознавание между двумя классами, заданными двумя одномерными нормальными распределениями с неравными априорными вероятностями.

На рис. 4.2 рассмотренная ситуация несколько модифицирована – объекты класса 1 встречаются при случайном выборе в 4 раза чаще, чем объекты класса 2. Это приводит к тому. что граница разделения между классами сдвигается правее – примерно в точку X_1=5.

Другими словами, мы пользуемся байесовским правилом – относим наблюдение к классу, имеющемму наибольшую апостериорную вероятность.

В рассмотренных двух примерах мы предположили, что известны теоретические распределения, определяющие классы. Если имеются только выборки наблюдений из них, то по ним можно построить оценки распределений. В частности, если есть основания считать, что неизвестные распределения нормальны, то достаточно вычислить средние и дисперсии двух выборок. Априорные вероятности можно взять равными долям встречаемости наблюдений двух классов в выборке.

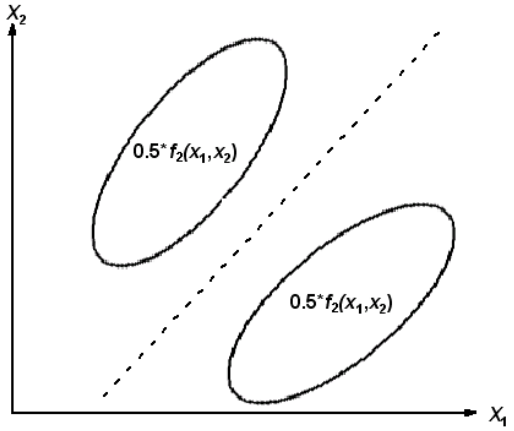


Рис. 4.3 Распознавание между двумя классами, заданными двумя двумерными нормальными распределениями с равными априорными вероятностями.

На рис. 4.3 показана ситуация, когда классы заданы двумя двумерными нормальными распределениями с равными дисперсиями и корреляциями и одинаковыми априорными вероятностями, отличаясь, однако, своими математическими ожиданиями. Каждая плотность схематично представлена одной из своих линий уровня (взято одно и тоже значение уровня для обеих плотностей). Легко показать, что в этом случае линия разделения между двумя классами, т.е. линия, вдоль которой значения обеих плотностей равны, является прямой (заметим, кстати, что в данном примере на одномерных проекциях на оси X_1 и X_2 распределения разделялись бы гораздо хуже, чем в двумерном пространстве). Действительно, условие равенства плотностей имеет вид

$$\frac{pk_1 \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1^{(1)})^2}{\sigma_1^2}+\frac{(x_2-\mu_2^{(1)})^2}{\sigma_2^2}-2\rho\frac{(x_1-\mu_1^{(1)})(x_2-\mu_2^{(1)})}{\sigma_1\sigma_2}\right]\right\}}{qk_1 \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1^{(2)})^2}{\sigma_1^2}+\frac{(x_2-\mu_2^{(2)})^2}{\sigma_2^2}-2\rho\frac{(x_1-\mu_1^{(2)})(x_2-\mu_2^{(2)})}{\sigma_1\sigma_2}\right]\right\}}=1$$

где $(\mu_1^{(1)}, \mu_2^{(1)})$ и $(\mu_1^{(2)}, \mu_2^{(2)})$ – координаты центров первого и второго распределений, σ_1 и σ_2 - их дисперсии по первой и второй оси, ρ - коэффициент корреляции, а коэффициенты k_1 определяется выражением

$$k_1 = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$

Сокращая на k_1 и логарифмируя, получаем соотношение

$$\left\{\log p-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1^{(1)})^2}{\sigma_1^2}+\frac{(x_2-\mu_2^{(1)})^2}{\sigma_2^2}-2\rho\frac{(x_1-\mu_1^{(1)})(x_2-\mu_2^{(1)})}{\sigma_1\sigma_2}\right]\right\}$$
$$-\left\{\log q-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1^{(2)})^2}{\sigma_1^2}+\frac{(x_2-\mu_2^{(2)})^2}{\sigma_2^2}-2\rho\frac{(x_1-\mu_1^{(2)})(x_2-\mu_2^{(2)})}{\sigma_1\sigma_2}\right]\right\} \quad (4.1)$$
$$=0$$

из которого видно, что после возведений в квадрат и перемножений выражений в скобках члены второго порядка x_1^2, x_2^2 и $x_1 x_2$ сократятся и в левой части уравнения останется выражение первого порядка относительно x_1 и x_2, т.е. это уравнение, действительно, определяет прямую на плоскости (x_1, x_2). Заметим, что этот вывод существенно использует предположение о равенстве соответствующих дисперсий и коэффициентов корреляций двух распределений, что эквивалентно равенству их ковариационных матриц C_1 и C_2

$$C_1 = C_2 = C = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

На рис. 4.4 показан случай, когда надо разделить три класса. Здесь также предполагается, что соответствующие дисперсии и корреляции трех распределений равны, поэтому каждая пара разделяется прямой равной плотности. Совокупность этих прямых формирует, как видно из рисунка, результирующую кусочно-линейную границу разделения классов на плоскости.

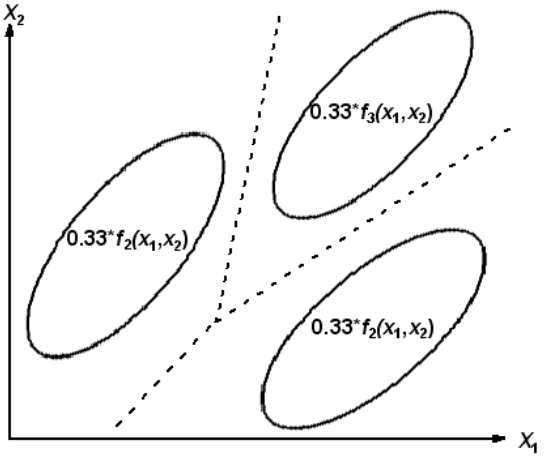


Рис. 4.4. Распознавание между тремя классами, заданными тремя двумерными нормальными распределениями с равными априорными вероятностями.

В предположении равенства ковариационных матриц и многомерной нормальности всех распределений рассмотренная ситуация обобщается и на большее число измерений. В трехмерном пространстве классы будут разделяться плоскостями, в четырехмерном – трехмерными гиперплоскостями и т.д. Именно предположение о нормальности и равенстве ковариационных матриц распределений, соответствующих разным классам, определяет условия применимости дискриминантного анализа, точнее, линейного дискриминантного анализа. В принципе, эти условия довольно жестки, однако, благодаря своей статистической прозрачности и легкости численной реализации, метод линейного дискриминантного анализа реализован во всех пакетах программ статистической обработки данных и широко применяется при анализе данных даже в ситуациях, когда его предпосылки выполняются лишь приближенно.

2. Линейная дискриминантная функция.

Рассмотрим еще некоторые понятия, связанные с линейным дискриминантным анализом.

Левую часть равенства (4.1), которую после преобразований можно записать в виде

$$L_{12}(x_1, x_2) = a_{12,0} + a_{12,1}x_1 + a_{12,2}x_2$$

называют линейной дискриминантной функцией классов для классов 1 и 2. Если для некоторого нового наблюдения эта функция больше нуля, то наблюдение следует отнести к классу 1, а если меньше – к классу 2.

Очевидно, для случая K классов имеется K(K-1)/2 дискриминантных функций и все их надо сравнить с нулем. В случае большого числа классов процедура становится слишком громоздкой, поэтому поступают иначе. С каждым классом k связывают функцию $L_k(x_1,x_2)$, называемую линейным дискриминантом этого класса, и для нового наблюдения вычисляют значения дискриминантов для всех классов. Полученные K значений сравниваются и наблюдение относится к тому классу, которому соответствует максимальное значение дискриминанта. Если вернуться к уравнению (3.29), то легко видеть, что его левую часть можно представить как разность двух линейных дискриминантов, соответствующим первому и второму членам в левой части (после раскрытия скобок и взаимного уничтожения квадратичных членов).

В пространстве наблюдаемых переменных для любого наблюдения можно вычислить его евклидовы расстояния от центра каждого класса. Однако эти расстояния могут не отражать истинные близости наблюдения к разным классам. Например, на рис. 4.2 точка на пересечении трех разделяющих прямых находится на разном расстоянии от центров трех классов, однако с точки зрения ее отнесения к одному из классов она равноудалена от них. Адекватной мерой расстояния в ситуации дискриминантного анализа (при равных априорных расстояниях) является так называемое расстояние Махаланобиса. Расстояние Махаланобиса между двумя точками в пространстве наблюдаемых переменных определяется как евклидово расстояние между этими точками в новом пространстве, получаемым из исходного пространства путем такого его преобразования, при котором первоначальные произвольные нормальные распределения преобразуются в нормальные распределения с единичными дисперсиями и нулевыми коэффициентами корреляциями. Например, в случае рис. 3.10 этого можно достичь путем поворота координатных осей в направлении осей эллипсов, представляющих линии уровня, и последующего соответствующего изменения масштаба новых осей.

Если предпосылки линейного дискриминантного анализа не выполняются даже приближенно, то используют нелинейные методы распознавания образов. Одним из таких методов является метод k ближайших соседей. Он состоит в том, что находятся k ближайших соседей нового наблюдения, и оно относится к тому классу, представителей которого больше всего среди этих ближайших соседей. Выбор числа k не формализован - алгоритм работает даже при k =1, но очевидно, что при большом числе наблюдений его можно брать большим, поскольку это должно повысить надежность правильного распознавания. Может оказаться полезным перед вычислением расстояний от нового наблюдения до всех остальных (для выявления ближайших) произвести преобразование пространства с тем, чтобы новые расстояния лучше соответствовали реальной близости наблюдений к своим классам. Во всяком случае, обязательно стоит пронормировать наблюдаемые переменные путем

деления их на свои выборочные среднеквадратичные отклонения, если эти переменные измерены в разных единицах.

3. Алгоритм дискриминантного анализа.

Основной целью дискриминации является нахождение такой линейной комбинации переменных (в дальнейшем эти переменные будем называть дискриминантными переменными), которая бы оптимально разделила рассматриваемые группы. Линейная функция

$$d_{pm} = \beta_0 + \beta_1 x_{1pm} + \ldots \beta_{wpm} \quad m = 1,\ldots,N \quad p = 1,\ldots,g \qquad (4.2)$$

называется канонической дискриминантной функцией с неизвестными коэффициентами β_i. Здесь d_{pm} значение дискриминантной функции для m–го объекта в группе k; x_{wpm} значение дискриминантной переменной для m–го объекта в группе p. С геометрической точки зрения дискриминантные функции определяют гиперповерхности в p-мерном пространстве. В частном случае при p=2 она является прямой, а при p=3 — плоскостью.

Коэффициенты β_i первой канонической дискриминантной функции выбираются таким образом, чтобы центроиды различных групп как можно больше отличались друг от друга. Коэффициенты второй группы выбираются также, но при этом налагается дополнительное условие, чтобы значения второй функции были некоррелированы со значениями первой. Аналогично определяются и другие функции. Отсюда следует, что любая каноническая дискриминантная функция d имеет нулевую внутригрупповую корреляцию с $d_1,\ldots,d_{g-1}$. Если число групп равно g, то число канонических дискриминантных функций будет на единицу меньше числа групп. Однако по многим причинам практического характера полезно иметь одну, две или же три дискриминантных функций. Тогда графическое изображениее объектов будет представлено в одно–, двух– и трехмерных пространствах. Такое представление особенно полезно в случае, когда число дискриминантных переменных p велико по сравнению с числом групп g.

Для получения коэффициентов β_i канонической дискриминантной функции нужен статистический критерий различения групп. Очевидно, что классификация переменных будет осуществляться тем лучше, чем меньше рассеяние точек относительно центроида внутри группы и чем больше расстояние между центроидами групп. Разумеется, что большая внутригрупповая вариация нежелательна, так как в этом случае любое заданное расстояние между двумя средними тем менее значимо в статистическом смысле, чем больше вариация распределений, соответствующих этим средним. Один из методов поиска наилучшей дискриминации данных заключается в нахождении такой канонической дискриминантной функции d, которая бы максимизировала отношение межгрупповой вариации к внутригрупповой

$$\lambda = B(d)/W(d)$$

где B — межгрупповая и W внутригрупповая матрицы рассеяния наблюдаемых переменных от средних. Рассмотрим максимизацию отношения (2) для произвольного числа классов. Введем следующие обозначения:

g — число классов;

p - число дискриминантных переменных;

N — число наблюдений в k–й группе;

N — общее число наблюдений по всем группам;

x_{ipm} — величина переменной i для m–го наблюдения в p–й группе;

x_{ip} — средняя величина переменной i в p–й группе;

$\bar{x}_i$ — среднее значение переменной i по всем группам;

T — общая сумма перекрестных произведений для переменных u и v;

W — внутригрупповая сумма перекрестных произведений для переменных u и v.; .

В модели дискриминации должны соблюдаться следующие условия:

1. число групп: $g \geq 2$;
2. число объектов в каждой группе: $n_i \geq 2$;
3. число дискриминантных переменных: $0 < p < (n-2)$;
4. дискриминантные переменные измеряются в интервальной шкале;
5. дискриминантные переменные линейно независимы;
6. ковариационные матрицы групп примерно равны;
7. дискриминантные переменные в каждой группе подчиняются многомерному нормальному закону распределения.

Рассмотрим задачу максимизации отношения (2) когда имеются g групп. Оценим сначала информацию, характеризующую степень различия между объектами по всему пространству точек, определяемому переменными групп. Для этого вычислим матрицу рассеяния T, которая равна сумме квадратов отклонений и попарных произведений наблюдений от общих средних $\bar{x}_i, i = 1,\ldots, p$ по каждой переменной. Элементы матрицы T определяются выражением

$$t_{ij} = \sum_{k=1}^{g}\sum_{m=1}^{n_k}(x_{ikm} - \bar{x}_i)(x_{jkm} - \bar{x}_j), \quad (4.3)$$

где; $\bar{x}_i = (1/N)\sum_{k=1}^{g} n_i \bar{x}_{ik},\ i = 1,\ldots, p \qquad \bar{x}_{ik} = (1/N_i)\sum_{m=1}^{N_k} x_{ikm},\ i = 1,\ldots, p; k = 1,\ldots, g$

Запишем это выражение в матричной форме. Обозначим p–мерную случайную векторную переменную k-ой группы следующим образом.

$$X_k = \{x_{ikm}\}, i = 1,\ldots, p,\ k = 1,\ldots, g,\ m = 1,\ldots, N_k$$

Тогда объединенная p-мерная случайная векторная переменная всех групп будет иметь вид

Общее среднее этой p-мерной случайной векторной переменной будет равен вектору средних отдельных признаков

$$X = [X_1 X_2 \ldots X_g]$$

Матрица рассеяния от среднего при этом запишется в виде

$$T = \sum_{k=1}^{g} (X_k - \bar{x})(X_k - \bar{x})' \quad (4)$$

Если использовать векторную переменную объединенных переменных X, то матрица T определится по формуле

$$T = (X - \bar{x})(X - \bar{x})'$$

Матрица T содержит полную информацию о распределении точек по пространству переменных. Диагональные элементы представляют собой сумму квадратов отклонений от общего среднего и показывают как ведут себя наблюдения по отдельно взятой переменной. Внедиагональные элементы равны сумме произведений отклонений по одной переменной на отклонения по другой.

Если разделить матрицу T на (n-1) , то получим ковариационную матрицу. Для проверки условия линейной независимости переменных полезно рассмотреть вместо T нормированную корреляционную матрицу.

Для измерения степени разброса объектов внутри групп рассмотрим матрицу W, которая отличается от T только тем, что ее элементы определяются векторами средних для отдельных групп, а не вектором средних для общих данных. Элементы внутригруппового рассеяния определятся выражением

$$w_{ij} = \sum_{k=1}^{g} \sum_{m=1}^{N_k} (x_{ikm} - \bar{x}_{ik})(x_{jkm} - \bar{x}_{jk})$$

Запишем это выражение в матричной форме. Данным g групп будут соответствовать векторы средних

$$\bar{x}_1 = [\bar{x}_{11}\bar{x}_{21}\ldots\bar{x}_{p1}],$$

$$\ldots$$

$$\bar{x}_g = [\bar{x}_{1g}\bar{x}_{2g}\ldots\bar{x}_{pg}].$$

Тогда матрица внутригрупповых вариаций запишется в виде

$$W = \sum_{k=1}^{g} (X_k - \bar{x}_k)(X_k - \bar{x}_k)' \quad (4.5)$$

Если разделить каждый элемент матрицы W на (n–g), то получим оценку ковариационной матрицы внутригрупповых данных.

Когда центроиды различных групп совпадают, то элементы матриц T и W будут равны. Если же центроиды групп различные, то разница

$$B = T - W \quad (4.6)$$

будет определять межгрупповую сумму квадратов отклонений и попарных произведений. Если расположение групп в пространстве различается (т.е. их центроиды не совпадают), то степень разброса наблюдений внутри групп будет меньше межгруппового разброса. Отметим, что элементы матрицы B можно вычислить и по данным средних.

$$b_{ij} = \sum_{k=1}^{g} N_k (\bar{x}_{ik} - \bar{x}_i)(\bar{x}_{jk} - \bar{x}_j), \quad i, j = 1,\ldots, p \quad (4.7)$$

Матрицы W и B содержат всю основную информацию о зависимости внутри групп и между группами. Для лучшего разделения наблюдений на группы нужно подобрать коэффициенты дискриминантной функции из условия

максимизации отношения межгрупповой матрицы рассеяния к внутригрупповой матрице рассеяния при условии ортогональности дискриминантных плоскостей. Тогда нахождение коэффициентов дискриминантных функций сводится к решению задачи о собственных значениях и векторах. Это утверждение можно сформулировать так: если спроектировать групп p–мерных выборок на $(g-1)$ пространство, порожденное собственными векторами $(v_{1k},\ldots,v_{pk}), k=1,\ldots,g-1$, то отношение (2) будет максимальным, т.е. рассеивание между группами будет максимальным при заданном внутригрупповом рассеивании. Если бы мы захотели спроектировать g выборок на прямую при условии максимизации наибольшего рассеивания между группами, то следовало бы использовать собственный вектор $(v_{11},\ldots,v_{p1})$, соответствующий максимальному собственному числу λ_1 . При этом дискриминантные функции можно получать: по нестандартизованным и стандартизованным коэффициентам.

Пусть $\lambda_1 \geq \ldots \geq \lambda_p$ и $v_1,\ldots,v_p$ соответственно собственные значения и векторы. Тогда условие (4.4) в терминах собственных чисел и векторов запишется в виде

$$\lambda = \frac{\sum_k b_{jk} v_j v_k}{\sum_k w_{jk} v_j v_k},$$

что влечет равенство $\sum_k (b_{jk} - \lambda w_{jk}) v_k = 0$, или в матричной записи

$$(B - \lambda_i W) v_i = 0 \quad v_i' W v_j = \delta_{ij}, \tag{4.8}$$

где δ_{ij} — символ Кронекера. Таким образом, решение уравнения $|B - \lambda W| = 0$ позволяет нам определить компоненты собственных векторов, соответствующих дискриминантным функциям. Если B и W невырожденные матрицы, то собственные корни уравнения $|B - \lambda W| = 0$ такие же, как и у $|W^{-1}B - \lambda I| = 0$. Решение системы уравнений (10) можно получить путем использования разложения Холецкого LL' матрицы W^{-1} и решения задачи о собственных значениях.

$$(L'BL - \lambda_i I) v_i = 0 \quad v_i' v_j = \delta_{ij}$$

Каждое решение, которое имеет свое собственное значение λ_i и собственный вектор v_i , соответствует одной дискриминантной функции. Компоненты собственного вектора v_i можно использовать в качестве коэффициентов дискриминантной функции. Однако при таком подходе начало координат не будет совпадать с главным центроидом. Для того чтобы начало координат совпало с главным центроидом нужно нормировать компоненты собственного вектора

$$\beta_i = v_i \sqrt{N - g} \quad \beta_0 = -\sum_{i=1}^{p} \beta_i \bar{x}_i \tag{4.9}$$

Нормированные коэффициенты получены по нестандартизованным исходным данным, поэтому они называются нестандартизованными. Нормированные коэффициенты приводят к таким дискриминантным значениям, единицей измерения которых является стандартное квадратичное отклонение. При таком подходе каждая ось в преобразованном пространстве сжимается или растягивается таким образом, что соответствующее дискриминантное значение для данного объекта представляет собой число стандартных отклонений точки от главного центроида.

Стандартизованные коэффициенты можно получить двумя способами:

1. по формуле (4.9), если исходные данные были приведены к стандартной форме;
2. преобразованием нестандартизованных коэффициентов к стандартизованной форме:

$$c_i = \beta_i \sqrt{\frac{w_{ii}}{N-g}} \tag{4.10}$$

где w_{ii} сумма внутригрупповых квадратов i-й переменной.

Стандартизованные коэффициенты полезно применять для уменьшения размерности исходного признакового пространства переменных. Если абсолютная величина коэффициента для данной переменной для всех дискриминантных функций мала, то эту переменную можно исключить, тем самым сократив число переменных.

Структурные коэффициенты определяются коэффициентами взаимной корреляции между отдельными переменными и дискриминантной функцией. Если относительно некоторой переменной абсолютная величина коэффициента велика, то вся информация о дискриминантной функции заключена в этой переменной.

Структурные коэффициенты полезны при классификации групп. Структурный коэффициент можно вычислить и для переменной в пределах отдельно взятой группы. Тогда получаем внутригрупповой структурный коэффициент, который вычисляется по формуле:

$$s_{ij} = \sum_{k=1}^{p} r_{ik} c_{kj} = \sum_{k=1}^{p} \frac{w_{ik} c_{kj}}{\sqrt{w_{ii} w_{jj}}}, \tag{4.11}$$

где s_{ij} внутригрупповой структурный коэффициент для i–ой переменной и j–ой функции; r_{ik} — внутригрупповые структурные коэффициенты корреляции между переменными i и k; c_{kj} — стандартизованные коэффициенты канонической функции для переменной k и функции j.

Структурные коэффициенты по своей информативности несколько отличаются от стандартизованных коэффициентов. Стандартизованные коэффициенты показывают вклад переменных в значение дискриминантной функции. Если две переменные сильно коррелированы, то их стандартизованные коэффициенты могут быть меньше по сравнению с теми случаями, когда используется только одна из этих переменных. Такое распределение величины стандартизованного коэффициента объясняется тем,

что при их вычислении учитывается влияние всех переменных. Структурные же коэффициенты являются парными корреляциями и на них не влияют взаимные зависимости прочих переменных.

Общее число дискриминантных функций не превышает числа дискриминантных переменных и, по крайней мере, на 1 меньше числа групп. Степень разделения выборочных групп зависит от величины собственных чисел: чем больше собственное число, тем сильнее разделение. Наибольшей разделительной способностью обладает первая дискриминантная функция, соответствующая наибольшему собственному числу λ_1, вторая обеспечивает максимальное различение после первой и т.д. Различительную способность i–ой функции оценивают по относительной величине в процентах собственного числа λ_i от суммы всех λ.

Другой характеристикой, позволяющей оценить полезность дискриминантной функции, является коэффициент канонической корреляции r_i. Каноническая корреляция является мерой связи между двумя множествами переменных. Максимальная величина этого коэффициента равна 1. Будем считать, что группы составляют одно множество, а другое множество образуют дискриминантные переменные. Коэффициент канонической корреляции для i–ой дискриминантной функции определяется формулой:

$$r_i = \sqrt{\frac{\lambda_i}{1+\lambda_i}} \tag{4.12}$$

Чем больше величина r_i, тем лучше разделительная способность дискриминантной функции.

Остаточная дискриминация.

Так как дискриминантные функции находятся по выборочным данным, они нуждаются в проверке статистической значимости. Дискриминантные функции представляются аналогично главным компонентам. Поэтому для проверки этой значимости можно воспользоваться критерием, аналогичным дисперсионному критерию в методе главных компонент. Этот критерий оценивает остаточную дискриминантную способность, под которой понимается способность различать группы, если при этом исключить информацию, полученную с помощью ранее вычисленных функций. Если остаточная дискриминация мала, то не имеет смысла дальнейшее вычисление очередной дискриминантной функции. Полученная статистика носит название "Λ – статистика Уилкса" и вычисляется по формуле:

$$\Lambda = \prod_{i=k+1}^{g}(1/(1+\lambda_i)) \tag{4.13}$$

где k — число вычисленных функций. Чем меньше эта статистика, тем значимее соответствующая дискриминантная функция. Величина

$$\chi^2 = -\left[N-((p+g)/2)-1\right]\ln\Lambda_k, \quad k=0,1,\ldots,g-1 \tag{4.14}$$

имеет хи–квадрат распределение с $(p-k)(g-k-1)$ степенями свободы.

Вычисления проводим в следующем порядке:

1. Находим значение критерия χ^2 при k=0. Значимость критерия подтверждает существование различий между группами. Кроме того, это доказывает, что первая дискриминантная функция значима и имеет смысл ее вычислять.

2. Определяем первую дискриминантную функцию и проверяем значимость критерия при k=1. Если критерий значим, то вычисляем вторую дискриминантную функцию и продолжаем процесс до тех пор, пока не будет исчерпана вся значимая формация.

Вопросы по лекции 4

1. Опишите область применения дискриминантнтного анализа.
2. Дайте определение задачи классификации.
3. Решающее правило и обоснование выбора дискриминационной функции.
4. Опишите процесс классификации объектов при помощи функции расстояния.
5. Понятие и отличия применения стандартизированных и нестандартизированных коэффициентов.
6. Полный алгоритм дискриминантного анализа
7. Назначение кластерного анализа
8. Процедура проведения кластерного анализа
9. Интерпретация результатов кластерного анализа
10. Интерпретация результатов дискриминантного анализа.

Практикум по модулю 1

Практическое занятие №1 Статистическая обработка данных с применением Excel

По соответствующему варианту, используя MS Excel:

1. Построить ряды распределения (интервальный и дискретный вариационные ряды). Изобразить их графики.

2. Построить график накопительных частот — кумуляту.

3. Составить эмпирическую функцию распределения и изобразить ее графически.

4. Вычислить моду, медиану, выборочную среднюю, выборочную дисперсию, выборочное среднее квадратическое отклонение, коэффициент вариации, асимметрию, эксцесс.

5. Построить доверительные интервалы для истинного значения измеряемой величины и среднего квадратического отклонения генеральной совокупности.

x - наблюдаемый признак

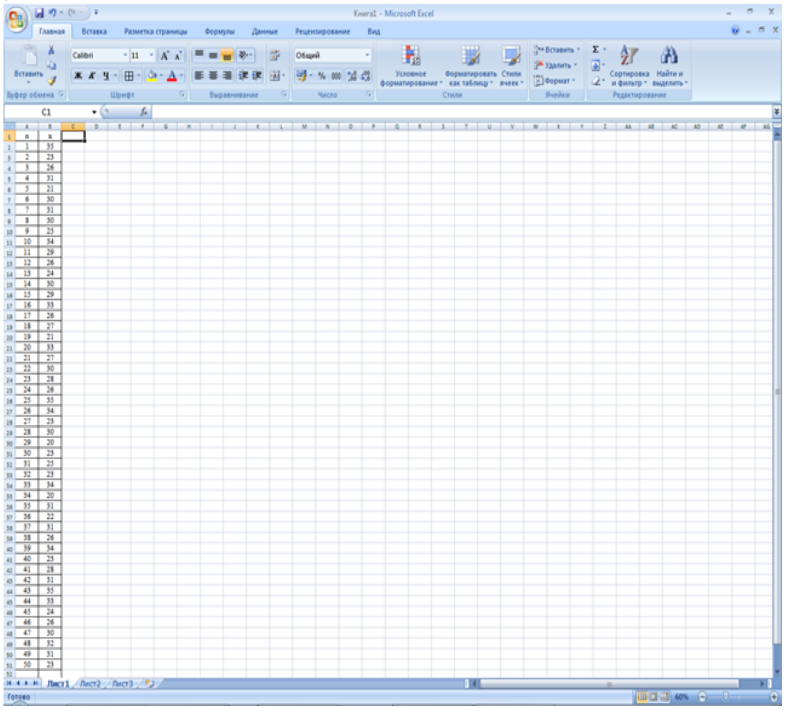
n - номер наблюдения

Пример решения задания

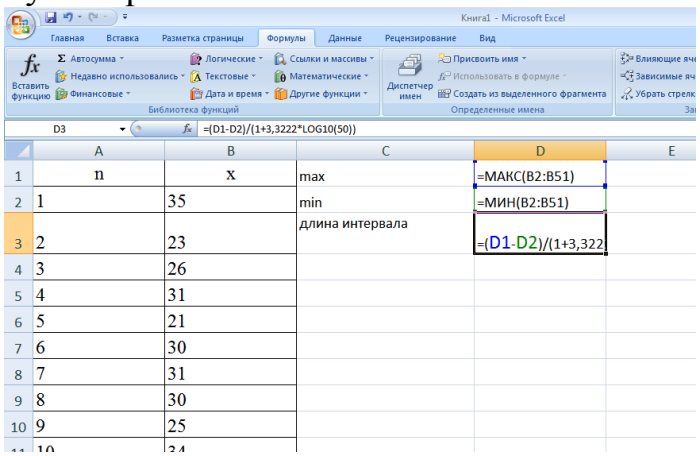
Исходные данные

n	1	2	3	4	5	6	7	8
x	35	23	26	31	21	30	31	30
n	9	10	11	12	13	14	15	16
x	25	34	29	26	24	30	29	33
n	17	18	19	20	21	22	23	24
x	26	27	21	33	27	30	28	26
n	25	26	27	28	29	30	31	32
x	35	34	23	30	20	23	25	23
n	33	34	35	36	37	38	39	40
x	34	20	31	22	31	26	34	23
n	41	42	43	44	45	46	47	48
x	28	31	35	33	24	26	30	32
n	49	50						
x	31	23						

1. Экспортируем данные в Excel

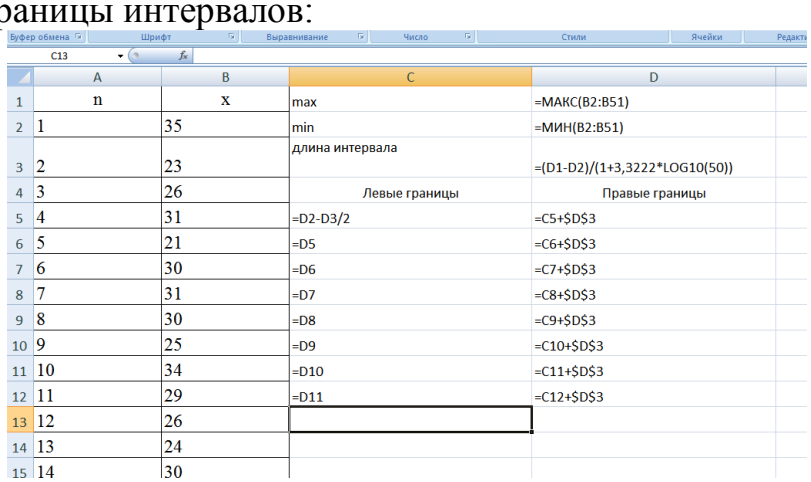


2. Используя мастер функций находим максимальное и минимальное значение в ряде и длину интервала



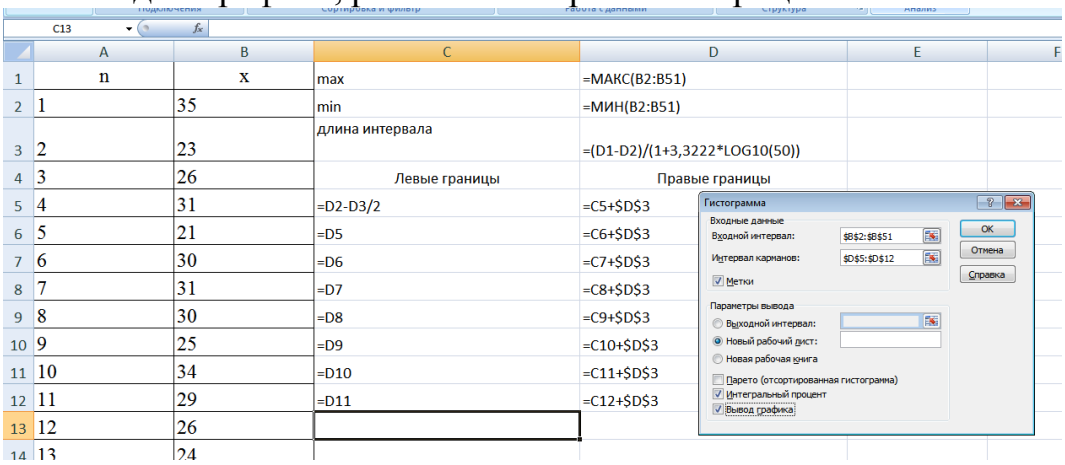
	A	B	C	D	E
1	n	x	max	=МАКС(B2:B51)	
2	1	35	min	=МИН(B2:B51)	
3	2	23	длина интервала	=(D1-D2)/(1+3,322	
4	3	26			
5	4	31			
6	5	21			
7	6	30			
8	7	31			
9	8	30			
10	9	25			

3. Строим границы интервалов:



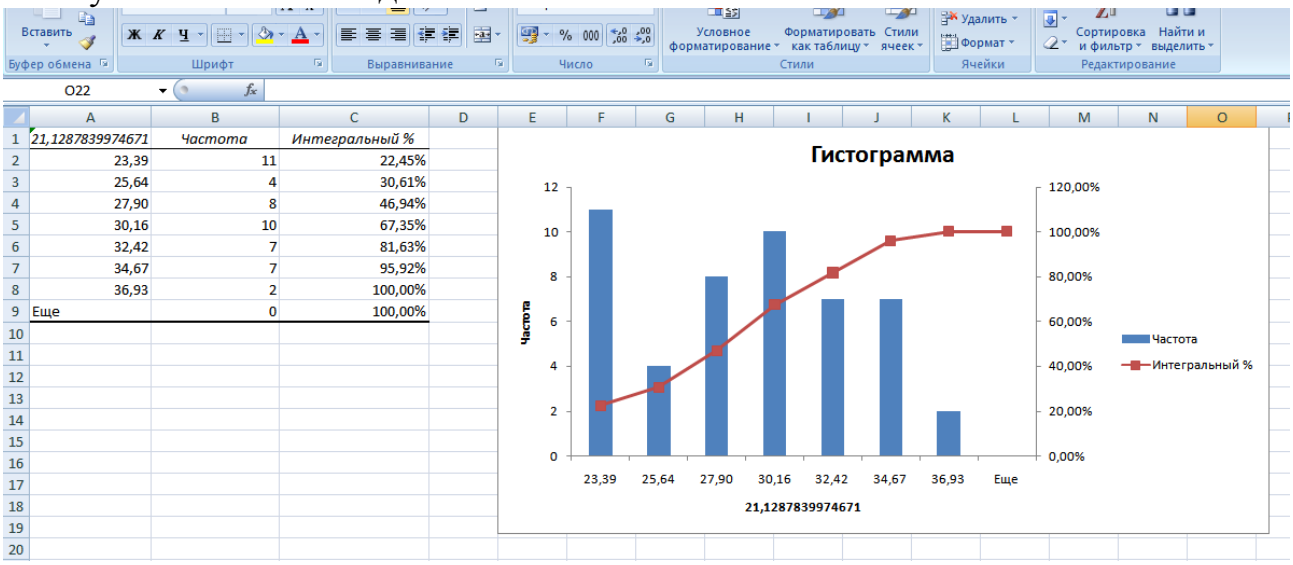
	A	B	C	D
1	n	x	max	=МАКС(B2:B51)
2	1	35	min	=МИН(B2:B51)
3	2	23	длина интервала	=(D1-D2)/(1+3,3222*LOG10(50))
4	3	26	Левые границы	Правые границы
5	4	31	=D2-D3/2	=C5+$D$3
6	5	21	=D5	=C6+$D$3
7	6	30	=D6	=C7+$D$3
8	7	31	=D7	=C8+$D$3
9	8	30	=D8	=C9+$D$3
10	9	25	=D9	=C10+$D$3
11	10	34	=D10	=C11+$D$3
12	11	29	=D11	=C12+$D$3
13	12	26		
14	13	24		
15	14	30		

4. Используя надстройку «Анализ данных» и вид анализа «Гистограма» строим с выводом графика, расчета интегральных процентов:



	A	B	C	D
1	n	x	max	=МАКС(B2:B51)
2	1	35	min	=МИН(B2:B51)
3	2	23	длина интервала	=(D1-D2)/(1+3,3222*LOG10(50))
4	3	26	Левые границы	Правые границы
5	4	31	=D2-D3/2	=C5+$D$3
6	5	21	=D5	=C6+$D$3
7	6	30	=D6	=C7+$D$3
8	7	31	=D7	=C8+$D$3
9	8	30	=D8	=C9+$D$3
10	9	25	=D9	=C10+$D$3
11	10	34	=D10	=C11+$D$3
12	11	29	=D11	=C12+$D$3
13	12	26		
14	13	24		

5. Результаты имеют вид:



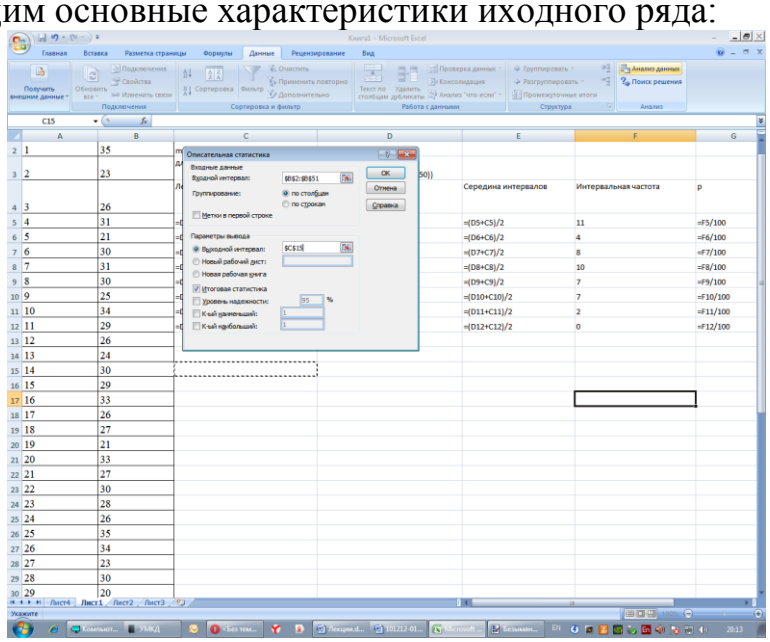
	A	B	C
1	*21,1287839974671*	*Частота*	*Интегральный %*
2	23,39	11	22,45%
3	25,64	4	30,61%
4	27,90	8	46,94%
5	30,16	10	67,35%
6	32,42	7	81,63%
7	34,67	7	95,92%
8	36,93	2	100,00%
9	Еще	0	100,00%

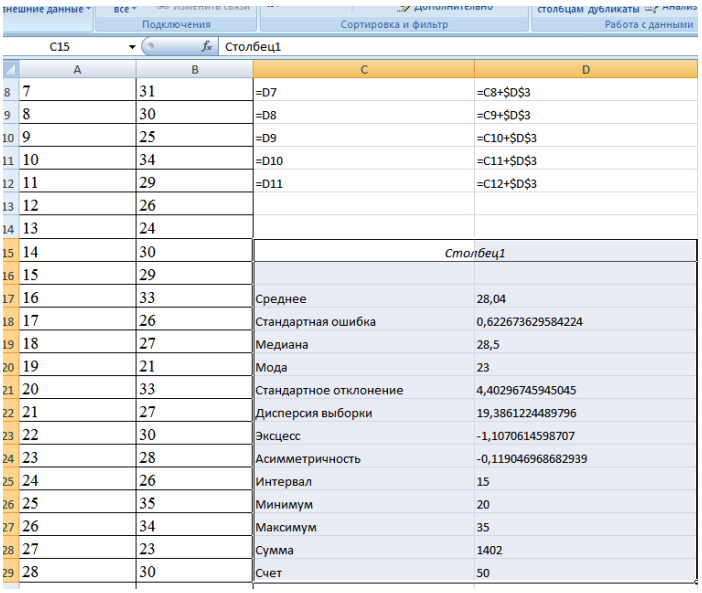
6. Рассчитаем дополнительные характеристики исходного ряда. Для этого введем следующие столбцы в таблицу Excel

n	x						
1	35	max	35				
2	23	min	20				
3	26	длина интервала	2,257567995				
4	31	Левые границы	Правые границы	Середина интервалов	Интервальная частота	p	Оценка функции плотности
5	21	18,871216	21,128784	20	11	0,11	0,048725
6	30	21,128784	23,38635199	22,25756799	4	0,04	0,017718182
7	31	23,38635199	25,64391999	24,51513599	8	0,08	0,035436363
8	30	25,64391999	27,90148798	26,77270398	10	0,1	0,044295454
9	25	27,90148798	30,15905598	29,03027198	7	0,07	0,031006818
10	34	30,15905598	32,41662397	31,28783997	7	0,07	0,031006818
11	29	32,41662397	34,67419197	33,54540797	2	0,02	0,008859091
12	26	34,67419197	36,93175996	35,80297596	0	0	0

max	=МАКС(B2:B51)				
min	=МИН(B2:B51)				
длина интервала	=(D1-D2)/(1+3,3222*LOG10(50))				
Левые границы	Правые границы	Середина интервалов	Интервальная частота	p	Оценка функции плотности
=D2-D3/2	=C5+$D$3	=(D5+C5)/2	11	=F5/100	=G5/$D$3
=D5	=C6+$D$3	=(D6+C6)/2	4	=F6/100	=G6/$D$3
=D6	=C7+$D$3	=(D7+C7)/2	8	=F7/100	=G7/$D$3
=D7	=C8+$D$3	=(D8+C8)/2	10	=F8/100	=G8/$D$3
=D8	=C9+$D$3	=(D9+C9)/2	7	=F9/100	=G9/$D$3
=D9	=C10+$D$3	=(D10+C10)/2	7	=F10/100	=G10/$D$3
=D10	=C11+$D$3	=(D11+C11)/2	2	=F11/100	=G11/$D$3
=D11	=C12+$D$3	=(D12+C12)/2	0	=F12/100	=G12/$D$3

7. Используя надстройку «Анализ данных:Описательная статистика» находим основные характеристики иходного ряда:





	A	B	C	D
8	7	31	=D7	=C8+$D$3
9	8	30	=D8	=C9+$D$3
10	9	25	=D9	=C10+$D$3
11	10	34	=D10	=C11+$D$3
12	11	29	=D11	=C12+$D$3
13	12	26		
14	13	24		
15	14	30	*Столбец1*	
16	15	29		
17	16	33	Среднее	28,04
18	17	26	Стандартная ошибка	0,622673629584224
19	18	27	Медиана	28,5
20	19	21	Мода	23
21	20	33	Стандартное отклонение	4,40296745945045
22	21	27	Дисперсия выборки	19,3861224489796
23	22	30	Эксцесс	-1,1070614598707
24	23	28	Асимметричность	-0,119046968682939
25	24	26	Интервал	15
26	25	35	Минимум	20
27	26	34	Максимум	35
28	27	23	Сумма	1402
29	28	30	Счет	50

Вариант 1

n	1	2	3	4	5	6	7	8
x	70	51	66	56	72	61	80	63
n	9	10	11	12	13	14	15	16

x	80	51	57	59	71	61	59	73
n	17	18	19	20	21	22	23	24
x	59	79	77	78	76	54	52	56
n	25	26	27	28	29	30	31	32
x	68	80	50	71	60	65	56	74
n	33	34	35	36	37	38	39	40
x	62	56	79	64	53	54	62	72
n	41	42	43	44	45	46	47	48
x	68	61	69	79	65	66	51	57
n	49	50						
x	73	65						

Вариант 2

n	1	2	3	4	5	6	7	8
x	70	27	12	22	20	21	62	47
n	9	10	11	12	13	14	15	16
x	35	62	39	66	57	50	13	35
n	17	18	19	20	21	22	23	24
x	35	46	36	25	70	37	47	27
n	25	26	27	28	29	30	31	32
x	19	34	16	64	28	56	60	48
n	33	34	35	36	37	38	39	40
x	53	66	36	18	65	26	54	22
n	41	42	43	44	45	46	47	48
x	34	11	15	46	59	69	38	14
n	49	50						
x	38	14						

Вариант 3

n	1	2	3	4	5	6	7	8
x	32	49	32	46	34	34	35	53
n	9	10	11	12	13	14	15	16
x	41	35	48	47	50	48	50	50
n	17	18	19	20	21	22	23	24
x	56	56	41	55	37	57	36	37
n	25	26	27	28	29	30	31	32
x	59	53	54	43	50	60	59	37
n	33	34	35	36	37	38	39	40
x	31	30	32	45	36	52	58	46
n	41	42	43	44	45	46	47	48
x	59	34	43	39	52	38	57	59
n	49	50						
x	57	57						

Вариант 4

n	1	2	3	4	5	6	7	8
x	39	38	40	32	38	33	37	39
n	9	10	11	12	13	14	15	16
x	39	37	35	39	33	34	37	39
n	17	18	19	20	21	22	23	24
x	38	37	33	31	36	33	36	34
n	25	26	27	28	29	30	31	32
x	33	33	37	31	35	34	39	35
n	33	34	35	36	37	38	39	40
x	39	33	38	34	32	38	40	32
n	41	42	43	44	45	46	47	48
x	39	30	35	33	39	32	31	32
n	49	50						
x	36	33						

Вариант 5

n	1	2	3	4	5	6	7	8
x	69	75	65	61	75	66	73	73
n	9	10	11	12	13	14	15	16
x	69	68	63	68	65	68	67	68
n	17	18	19	20	21	22	23	24
x	68	60	65	74	69	75	61	62
n	25	26	27	28	29	30	31	32
x	69	71	69	61	73	63	71	72
n	33	34	35	36	37	38	39	40
x	70	73	61	71	71	71	73	67
n	41	42	43	44	45	46	47	48
x	61	62	69	68	63	63	67	62
n	49	50						
x	65	67						

Вариант 6

n	1	2	3	4	5	6	7	8
x	70	75	87	82	77	88	81	74
n	9	10	11	12	13	14	15	16
x	72	85	74	72	86	72	86	74
n	17	18	19	20	21	22	23	24
x	74	81	80	85	73	72	81	86
n	25	26	27	28	29	30	31	32
x	79	85	82	79	77	75	85	83
n	33	34	35	36	37	38	39	40

x	83	90	81	78	72	87	76	74
n	41	42	43	44	45	46	47	48
x	80	77	83	70	81	71	79	84
n	49	50						
x	70	81						

Вариант 7

n	1	2	3	4	5	6	7	8
x	83	82	86	84	85	87	86	81
n	9	10	11	12	13	14	15	16
x	87	80	87	89	89	81	82	87
n	17	18	19	20	21	22	23	24
x	81	82	88	83	83	89	82	83
n	25	26	27	28	29	30	31	32
x	88	88	82	83	83	84	82	83
n	33	34	35	36	37	38	39	40
x	84	89	82	81	80	82	83	82
n	41	42	43	44	45	46	47	48
x	90	84	89	87	85	84	82	81
n	49	50						
x	86	85						

Вариант 8

n	1	2	3	4	5	6	7	8
x	54	56	55	51	57	54	59	59
n	9	10	11	12	13	14	15	16
x	56	59	56	56	58	52	54	57
n	17	18	19	20	21	22	23	24
x	57	53	56	53	59	55	51	51
n	25	26	27	28	29	30	31	32
x	58	50	57	53	54	56	59	53
n	33	34	35	36	37	38	39	40
x	60	57	58	54	60	52	56	56
n	41	42	43	44	45	46	47	48
x	57	57	53	52	53	52	57	50
n	49	50						
x	55	52						

Вариант 9

n	1	2	3	4	5	6	7	8
x	104	109	111	112	110	111	110	106
n	9	10	11	12	13	14	15	16
x	105	107	109	108	104	105	103	106
n	17	18	19	20	21	22	23	24

x	104	102	112	115	112	111	114	110
n	25	26	27	28	29	30	31	32
x	108	107	111	112	111	110	106	103
n	33	34	35	36	37	38	39	40
x	111	109	106	110	105	101	110	107
n	41	42	43	44	45	46	47	48
x	106	102	113	111	115	104	113	112
n	49	50						
x	104	110						

Вариант 10

n	1	2	3	4	5	6	7	8
x	35	23	26	31	21	30	31	30
n	9	10	11	12	13	14	15	16
x	25	34	29	26	24	30	29	33
n	17	18	19	20	21	22	23	24
x	26	27	21	33	27	30	28	26
n	25	26	27	28	29	30	31	32
x	35	34	23	30	20	23	25	23
n	33	34	35	36	37	38	39	40
x	34	20	31	22	31	26	34	23
n	41	42	43	44	45	46	47	48
x	28	31	35	33	24	26	30	32
n	49	50						
x	31	23						

Практическое занятие №2 Построение и анализ моделей парной, нелинейной, множественной регрессии

По выбранному варианту выполнить следующие задания:

1. Построить уравнение парной регрессии между фактором Y и фактором X1.

2. Построить уравнение нелинейной парной регрессии (логарифмическая, показательная или степенная функции на выбор аспиранта) между фактором Y и фактором X2

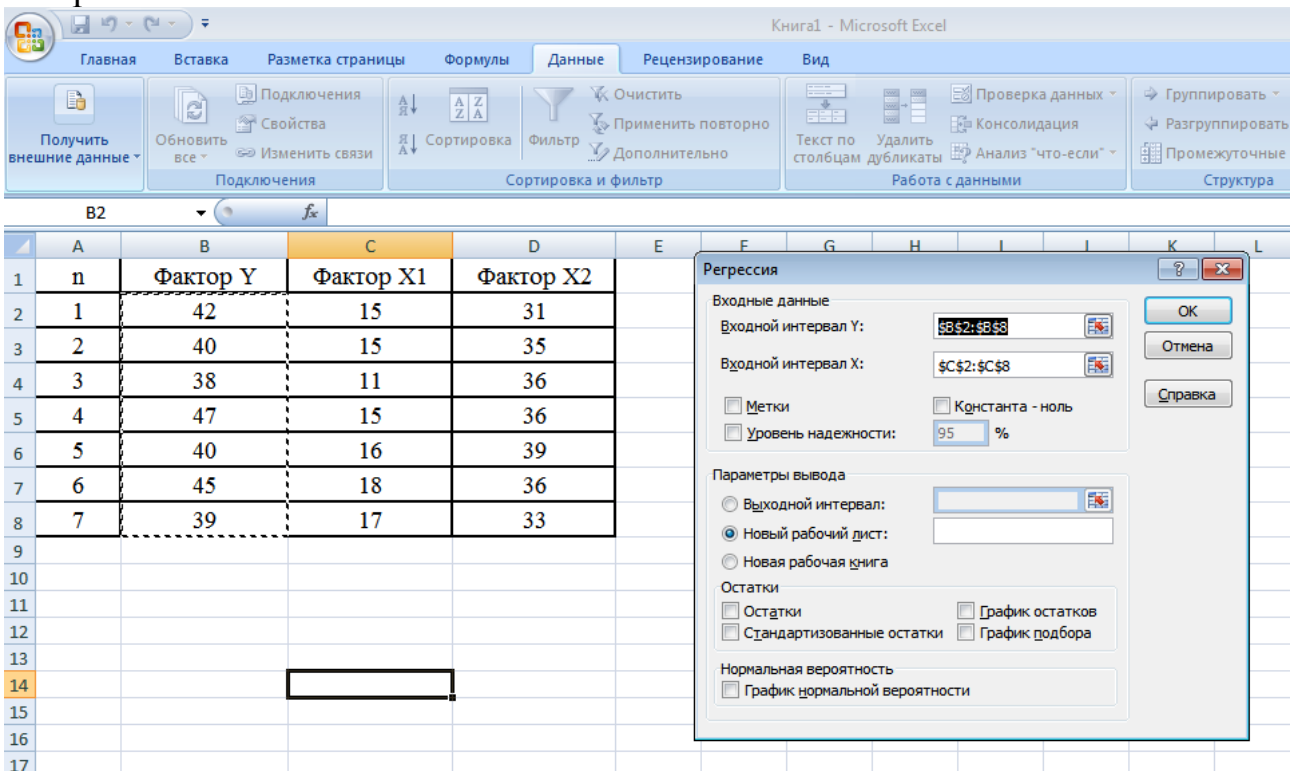
3. Построить линейное уравнение множественной регрессии.

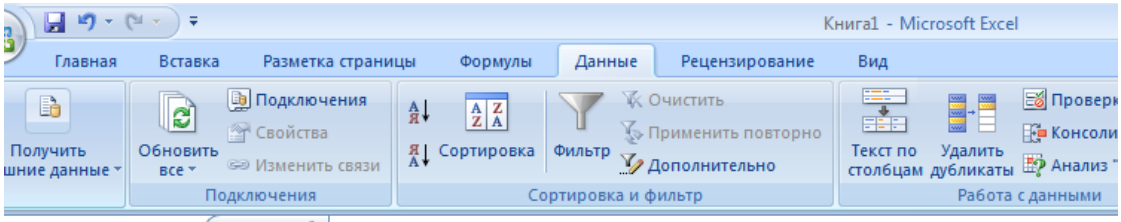
4. Для каждого этапа рассчитать коэффициент (индекс) корреляции.

Пример решения задачи

n	Фактор Y	Фактор X1	Фактор X2
1	42	15	31
2	40	15	35
3	38	11	36
4	47	15	36
5	40	16	39
6	45	18	36
7	39	17	33

1. Для построения уравнения парной регрессии воспользуемся надстройкой «Анализ данных: регрессия». Вводим соответствующие инетрвалы:





H29

	A	B	C	D	E	F	G	H	I	J
1	ВЫВОД ИТОГОВ									
2										
3	*Регрессионная статистика*									
4	Множественный R	0,406067								
5	R-квадрат	0,16489								
6	Нормированный R-квадрат	-0,00213								
7	Стандартная ошибка	3,312964								
8	Наблюдения	7								
9										
10	Дисперсионный анализ									
11		*df*	*SS*	*MS*	*F*	*ачимость F*				
12	Регрессия	1	10,83564494	10,83564	0,987237	0,366039				
13	Остаток	5	54,87864078	10,97573						
14	Итого	6	65,71428571							
15										
16		*Коэффициен*	*ндартная ошт*	*татисти*	*-Значени*	*ижние 95%*	*ерхние 95%*	*жние 95,0%*	*рхние 95,0%*	
17	Y-пересечение	32,29612	9,418682987	3,428942	0,018657	8,084621	56,50761	8,084621	56,50761	
18	Переменная X 1	0,606796	0,610705851	0,993598	0,366039	-0,96307	2,176665	-0,96307	2,176665	

Уравнение регрессии имеет вид: Y=32,29612+0,6068X1

Коэффициент корреляции равен 0,4061

2. Построим степенную функцию: $Y=aX2^b$

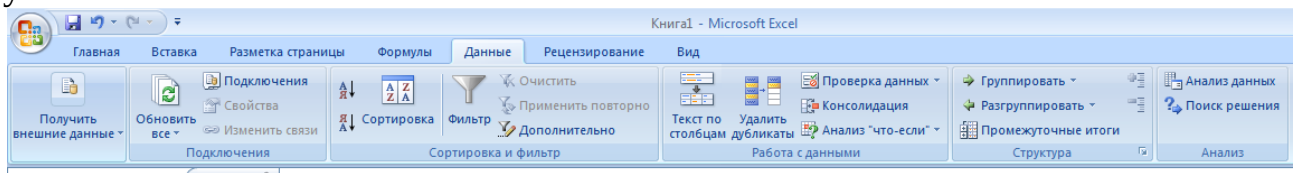
lnY=lna+blnX2

y=lnY

A=lna

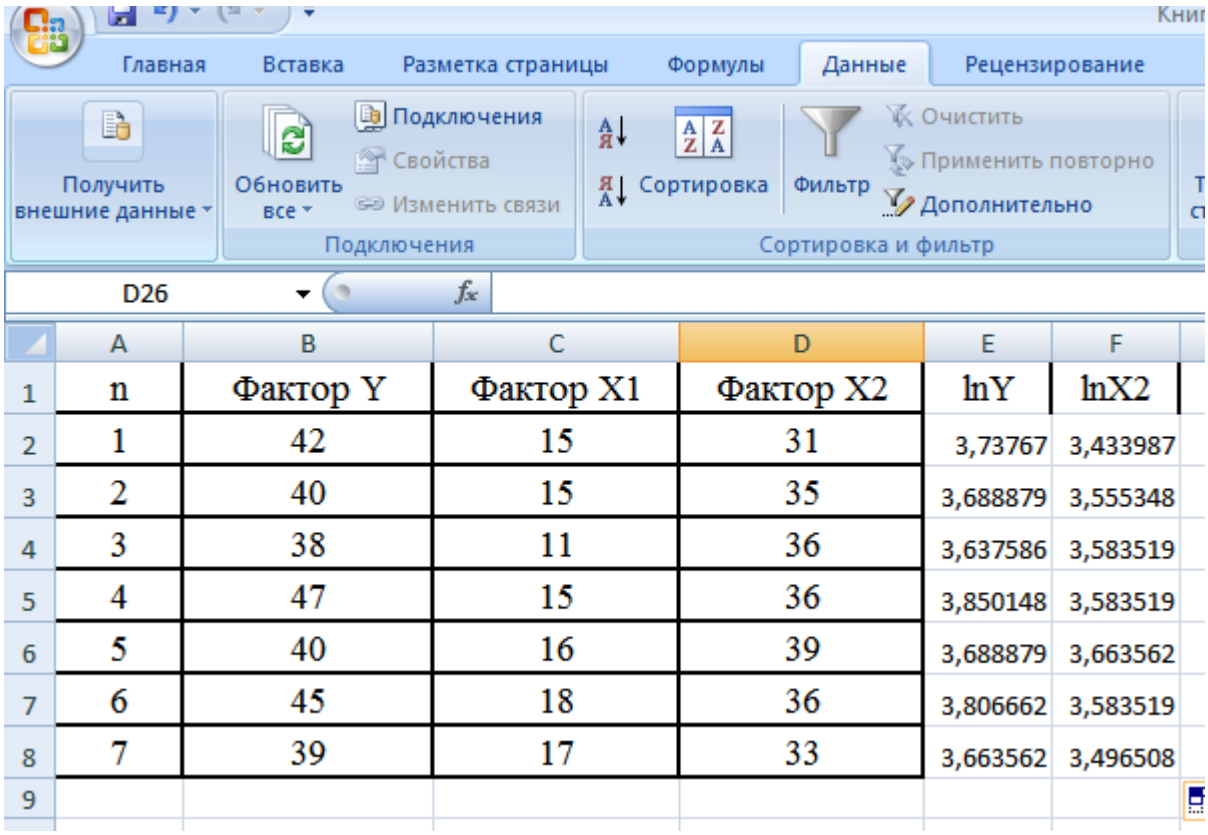
x2=lnX2

y=A+bx2



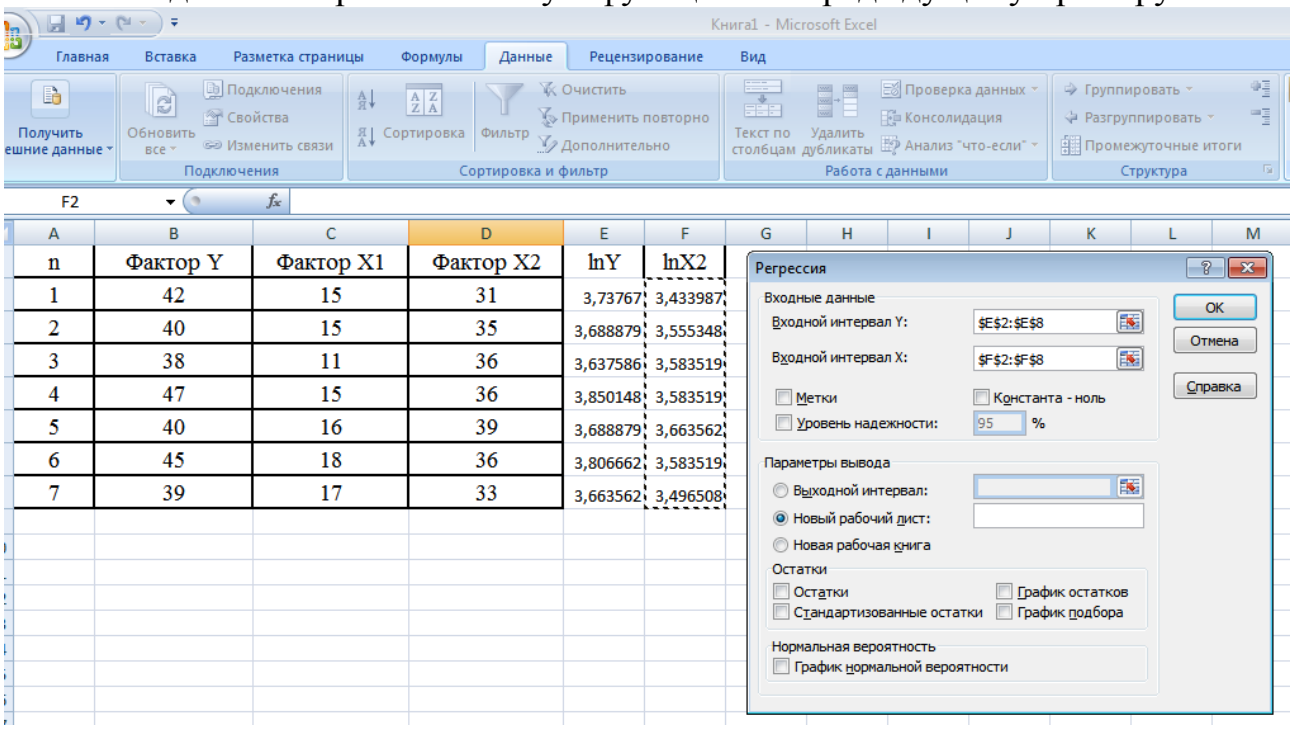
D26

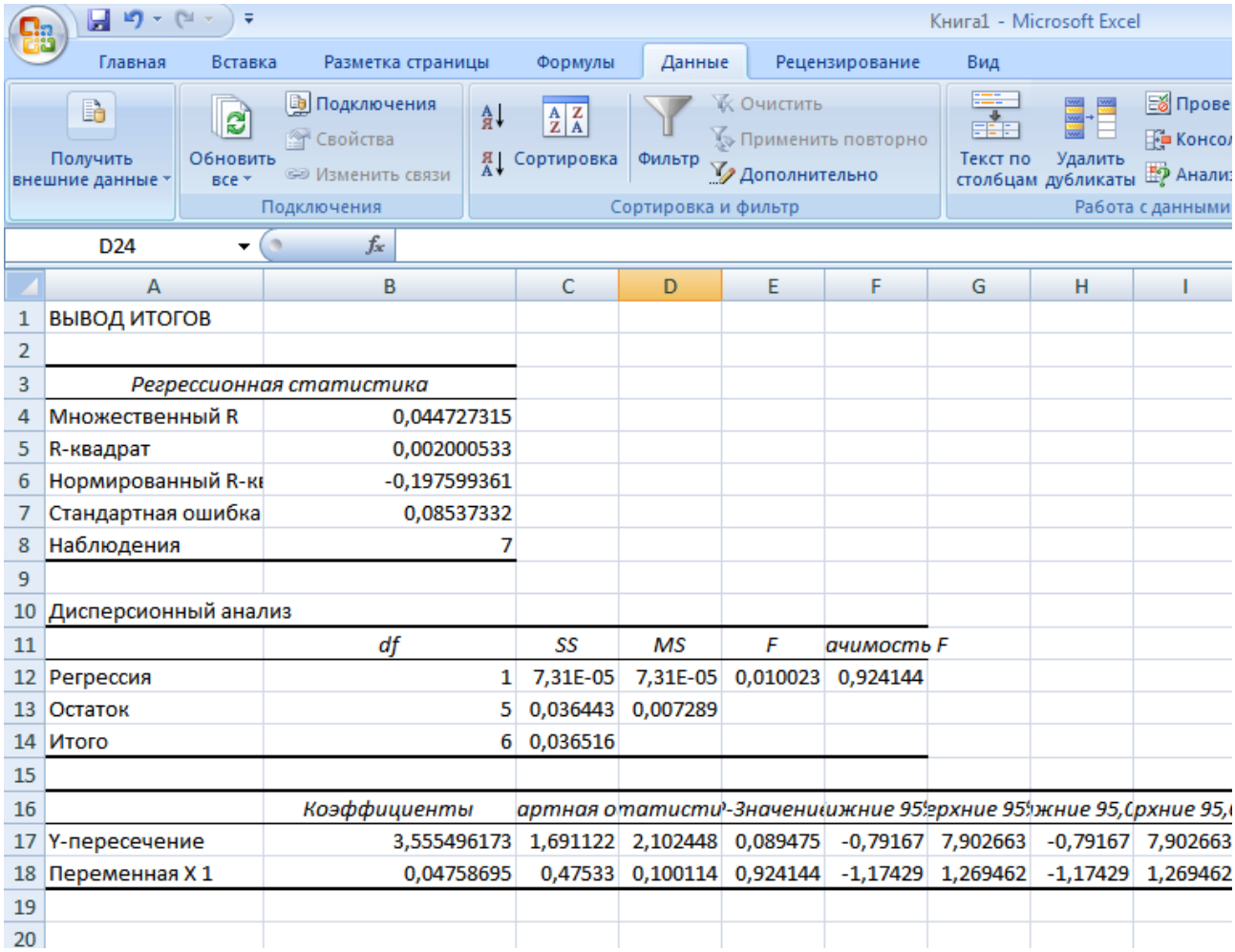
	A	B	C	D	E	F
1	n	Фактор Y	Фактор X1	Фактор X2	lnY	lnX2
2	1	42	15	31	=LN(B2)	=LN(D2)
3	2	40	15	35	=LN(B3)	=LN(D3)
4	3	38	11	36	=LN(B4)	=LN(D4)
5	4	47	15	36	=LN(B5)	=LN(D5)
6	5	40	16	39	=LN(B6)	=LN(D6)
7	6	45	18	36	=LN(B7)	=LN(D7)
8	7	39	17	33	=LN(B8)	=LN(D8)



	A	B	C	D	E	F
1	n	Фактор Y	Фактор X1	Фактор X2	lnY	lnX2
2	1	42	15	31	3,73767	3,433987
3	2	40	15	35	3,688879	3,555348
4	3	38	11	36	3,637586	3,583519
5	4	47	15	36	3,850148	3,583519
6	5	40	16	39	3,688879	3,663562
7	6	45	18	36	3,806662	3,583519
8	7	39	17	33	3,663562	3,496508
9						

По новым данным строим линейную функцию по предыдущему примеру





	A	B	C	D	E	F	G	H	I
1	ВЫВОД ИТОГОВ								
2									
3	*Регрессионная статистика*								
4	Множественный R	0,044727315							
5	R-квадрат	0,002000533							
6	Нормированный R-кв	-0,197599361							
7	Стандартная ошибка	0,08537332							
8	Наблюдения	7							
9									
10	Дисперсионный анализ								
11		*df*	*SS*	*MS*	*F*	*ачимость F*			
12	Регрессия	1	7,31E-05	7,31E-05	0,010023	0,924144			
13	Остаток	5	0,036443	0,007289					
14	Итого	6	0,036516						
15									
16		*Коэффициенты*	*артная о татисти*	*-Значени*	*ижние 95*	*рхние 95*	*жние 95,0*	*рхние 95,0*	
17	Y-пересечение	3,555496173	1,691122	2,102448	0,089475	-0,79167	7,902663	-0,79167	7,902663
18	Переменная X 1	0,04758695	0,47533	0,100114	0,924144	-1,17429	1,269462	-1,17429	1,269462
19									
20									

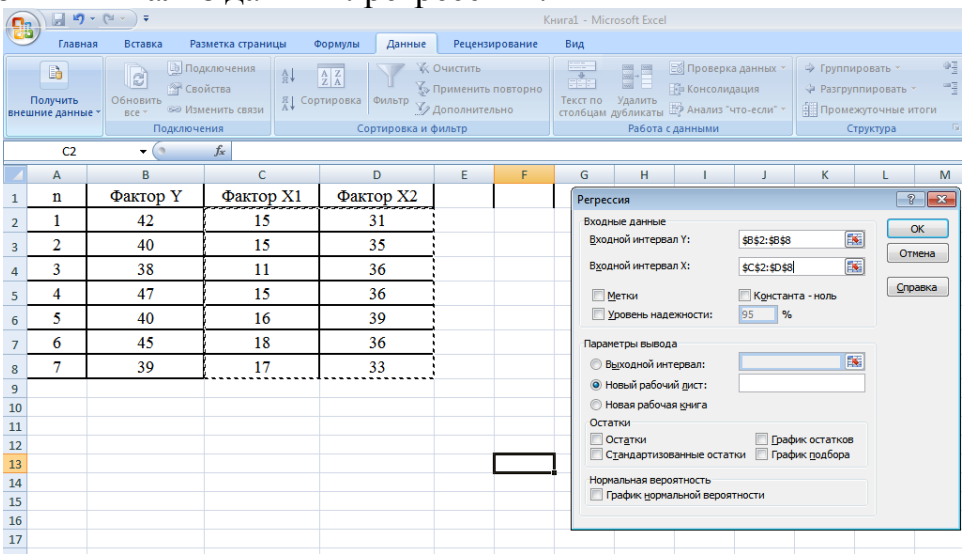
Уравнение имеет вид:

y=3,55+0,0476x2, тогда степенная функция получается посредством следующих обратных преобразованиq:

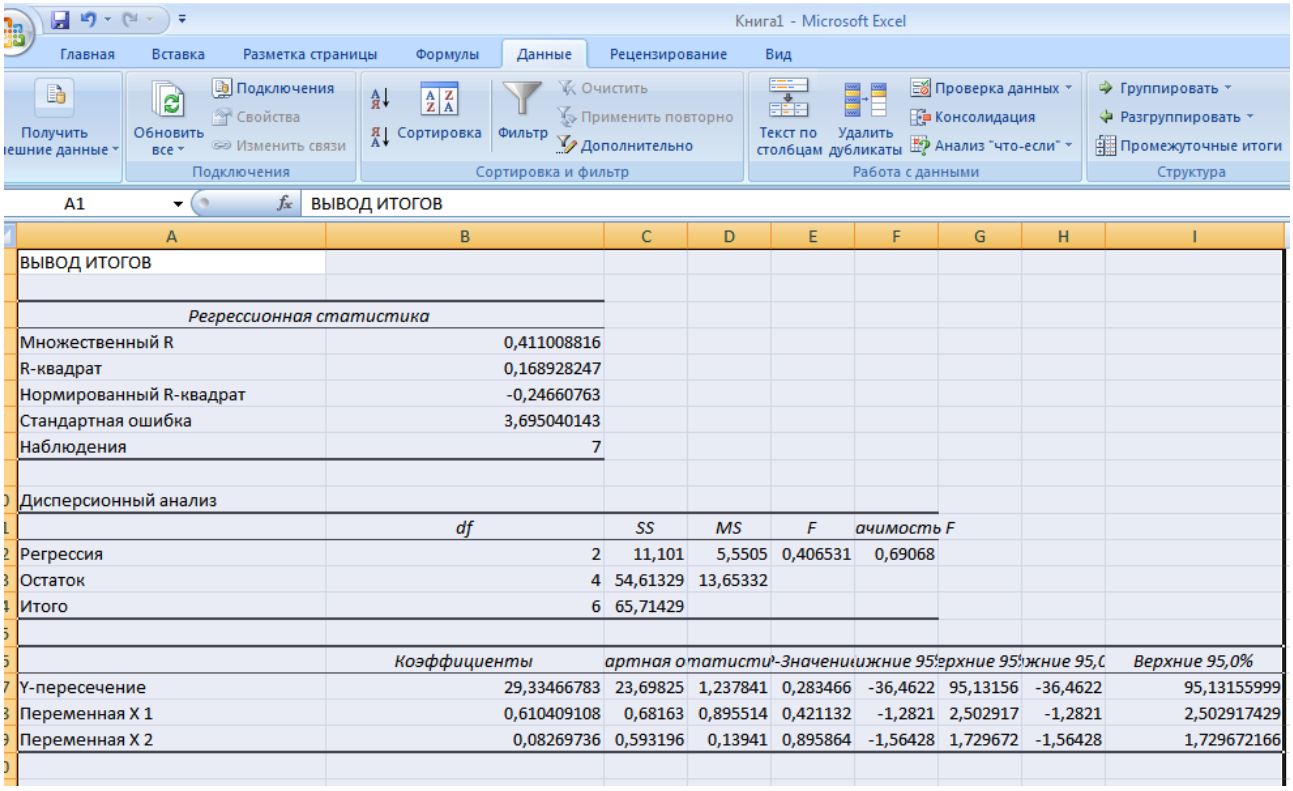
3,55=lna, $e^{3,55}$=a, a=35,005

$$Y=35,005*X2^{0,0476}$$

3. Построим уравнение множественной регрессии с использованием надстройки «Анализ данных: регрессия».



	A	B	C	D
1	n	Фактор Y	Фактор X1	Фактор X2
2	1	42	15	31
3	2	40	15	35
4	3	38	11	36
5	4	47	15	36
6	5	40	16	39
7	6	45	18	36
8	7	39	17	33



Книга1 - Microsoft Excel

Главная | Вставка | Разметка страницы | Формулы | Данные | Рецензирование | Вид

A1 | ВЫВОД ИТОГОВ

ВЫВОД ИТОГОВ								
Регрессионная статистика								
Множественный R	0,411008816							
R-квадрат	0,168928247							
Нормированный R-квадрат	-0,24660763							
Стандартная ошибка	3,695040143							
Наблюдения	7							
Дисперсионный анализ								
	df	*SS*	*MS*	*F*	*ачимость F*			
Регрессия	2	11,101	5,5505	0,406531	0,69068			
Остаток	4	54,61329	13,65332					
Итого	6	65,71429						
	Коэффициенты	*артная о*	*татисти*	*-Значени*	*ижние 95%*	*ерхние 95%*	*ижние 95,0*	*Верхние 95,0%*
Y-пересечение	29,33466783	23,69825	1,237841	0,283466	-36,4622	95,13156	-36,4622	95,13155999
Переменная X 1	0,610409108	0,68163	0,895514	0,421132	-1,2821	2,502917	-1,2821	2,502917429
Переменная X 2	0,08269736	0,593196	0,13941	0,895864	-1,56428	1,729672	-1,56428	1,729672166

Уравнение имеет вид:

$$Y=29,3347+0,61X1+0,083X2$$

Индекс корреляции равен =0,411

Варианты

Вариант 1

n	Фактор Y	Фактор X1	Фактор X2
1	38	15	31
2	32	15	35
3	32	11	36
4	39	15	36
5	32	16	39
6	37	18	36
7	31	17	33

Вариант 2

n	Фактор Y	Фактор X1	Фактор X2
1	37	32	40
2	36	33	35
3	38	34	19
4	27	32	32
5	37	33	31
6	31	30	36
7	20	32	43

Вариант 3

n	Фактор Y	Фактор X1	Фактор X2
1	28	115	51
2	34	111	64
3	39	101	76
4	30	100	79
5	25	116	67
6	30	111	80
7	29	107	79

Вариант 4

n	Фактор Y	Фактор X1	Фактор X2
1	79	24	101
2	68	21	99
3	77	13	94
4	70	10	103
5	82	28	94
6	61	19	102
7	74	29	97

Вариант 5

n	Фактор Y	Фактор X1	Фактор X2
1	43	27	53
2	37	22	64
3	39	23	81
4	39	23	78
5	41	30	84
6	37	22	72
7	49	22	64

Вариант 6

n	Фактор Y	Фактор X1	Фактор X2
1	83	24	23
2	54	54	39
3	55	26	28
4	79	43	30
5	61	21	29
6	92	56	35
7	66	57	20

Вариант 7

n	Фактор Y	Фактор X1	Фактор X2
1	14	27	93
2	18	24	80
3	16	12	64
4	17	27	85
5	16	9	77
6	26	5	74
7	21	12	87

Вариант 8

n	Фактор Y	Фактор X1	Фактор X2
1	42	134	86
2	29	101	84
3	38	114	100
4	40	113	101
5	39	149	85
6	48	135	77
7	24	119	91

Вариант 9

n	Фактор Y	Фактор X1	Фактор X2
1	29	56	110
2	19	37	109
3	28	60	121
4	26	34	125
5	19	36	102
6	25	69	126
7	27	69	119

Вариант 10

n	Фактор Y	Фактор X1	Фактор X2
1	20	45	11
2	40	40	18
3	55	34	25
4	63	31	32
5	81	36	39
6	95	29	46
7	109	26	53

Практическое занятие №3 Построение факторных моделей

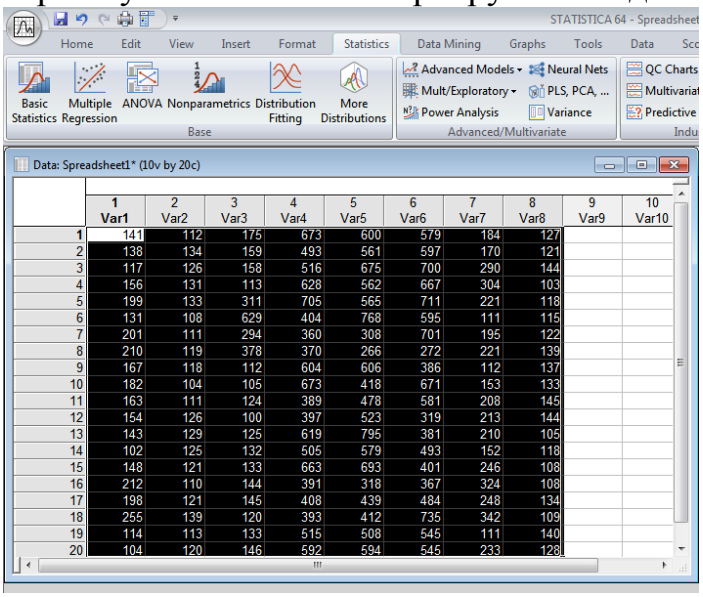
По предложенному варианту, используя ППП Statistica, SPSS или Deductor провести факторный анализ и дать интерпретацию полученным результатам.

X1, X2 … X8 некоторые признаки, характеризующие исследуемый процесс.

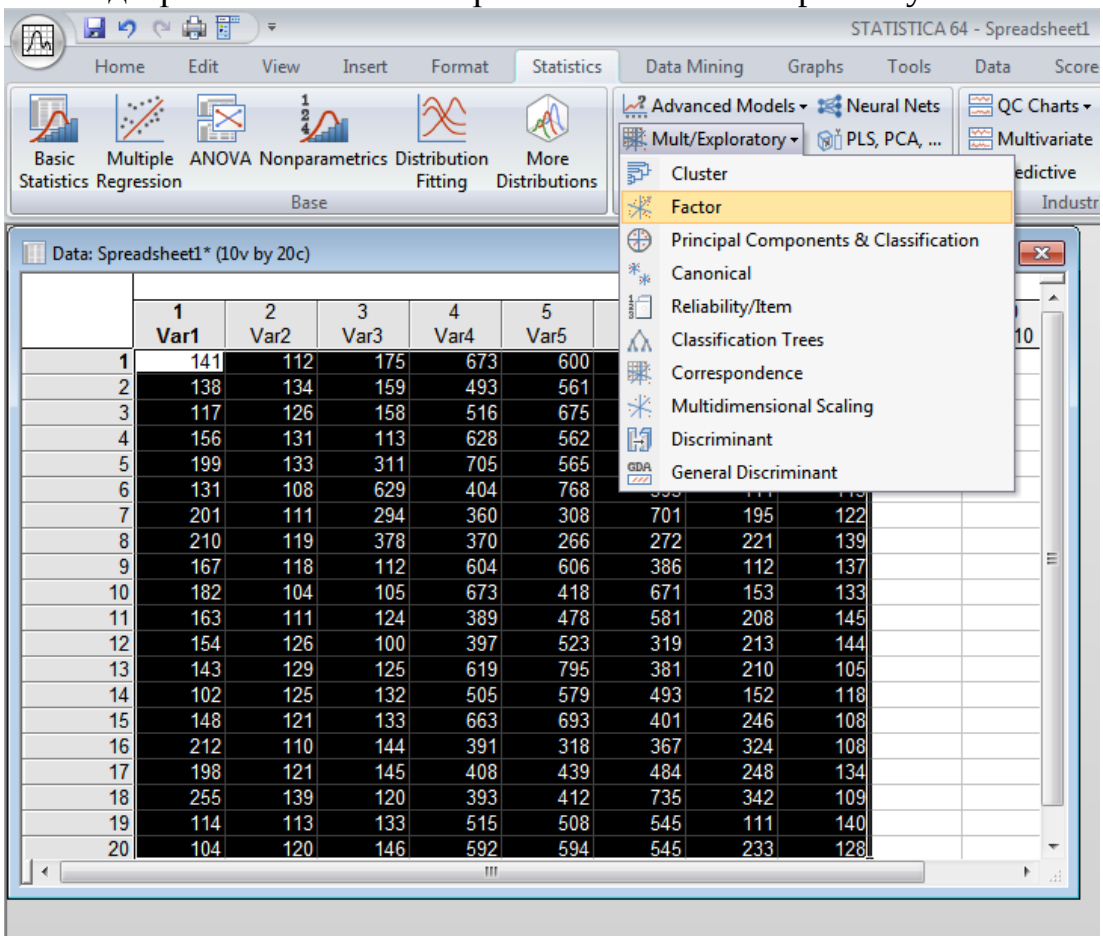
Пример решения задачи

n	X1	X2	X3	X4	X5	X6	X7	X8
1	141	112	175	673	600	579	184	127
2	138	134	159	493	561	597	170	121
3	117	126	158	516	675	700	290	144
4	156	131	113	628	562	667	304	103
5	199	133	311	705	565	711	221	118
6	131	108	629	404	768	595	111	115
7	201	111	294	360	308	701	195	122
8	210	119	378	370	266	272	221	139
9	167	118	112	604	606	386	112	137
10	182	104	105	673	418	671	153	133
11	163	111	124	389	478	581	208	145
12	154	126	100	397	523	319	213	144
13	143	129	125	619	795	381	210	105
14	102	125	132	505	579	493	152	118
15	148	121	133	663	693	401	246	108
16	212	110	144	391	318	367	324	108
17	198	121	145	408	439	484	248	134
18	255	139	120	393	412	735	342	109
19	114	113	133	515	508	545	111	140
20	104	120	146	592	594	545	233	128

1. Запускаем программу Statistica и экспортируем исходные данные:

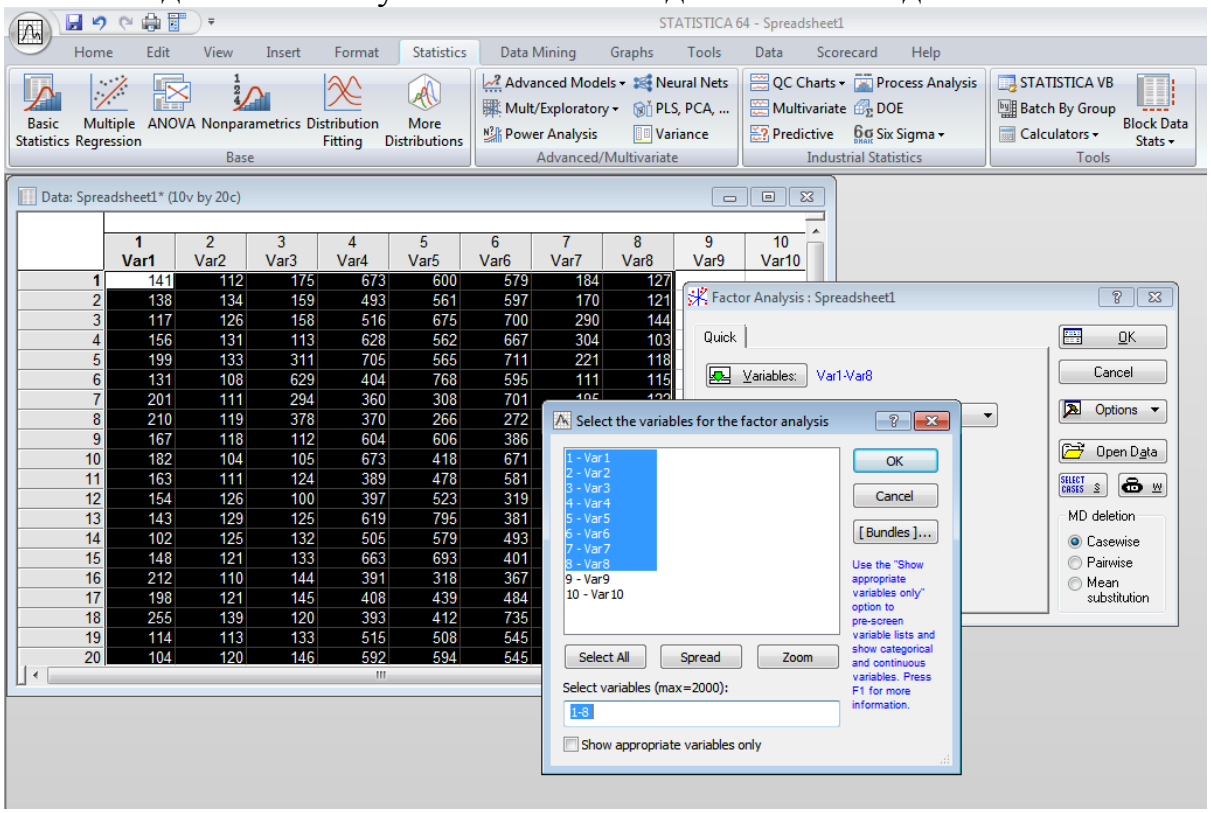


2. В надстройке Statistic выбираем анализ Mult/Exporatory - Factor:



	1 Var1	2 Var2	3 Var3	4 Var4	5 Var5			
1	141	112	175	673	600			
2	138	134	159	493	561			
3	117	126	158	516	675			
4	156	131	113	628	562			
5	199	133	311	705	565			
6	131	108	629	404	768			
7	201	111	294	360	308	701	195	122
8	210	119	378	370	266	272	221	139
9	167	118	112	604	606	386	112	137
10	182	104	105	673	418	671	153	133
11	163	111	124	389	478	581	208	145
12	154	126	100	397	523	319	213	144
13	143	129	125	619	795	381	210	105
14	102	125	132	505	579	493	152	118
15	148	121	133	663	693	401	246	108
16	212	110	144	391	318	367	324	108
17	198	121	145	408	439	484	248	134
18	255	139	120	393	412	735	342	109
19	114	113	133	515	508	545	111	140
20	104	120	146	592	594	545	233	128

3. На вкладке Variables указываем необходимый массив данных:



	1 Var1	2 Var2	3 Var3	4 Var4	5 Var5	6 Var6	7 Var7	8 Var8	9 Var9	10 Var10
1	141	112	175	673	600	579	184	127		
2	138	134	159	493	561	597	170	121		
3	117	126	158	516	675	700	290	144		
4	156	131	113	628	562	667	304	103		
5	199	133	311	705	565	711	221	118		
6	131	108	629	404	768	595	111	115		
7	201	111	294	360	308	701				
8	210	119	378	370	266	272				
9	167	118	112	604	606	386				
10	182	104	105	673	418	671				
11	163	111	124	389	478	581				
12	154	126	100	397	523	319				
13	143	129	125	619	795	381				
14	102	125	132	505	579	493				
15	148	121	133	663	693	401				
16	212	110	144	391	318	367				
17	198	121	145	408	439	484				
18	255	139	120	393	412	735				
19	114	113	133	515	508	545				
20	104	120	146	592	594	545				

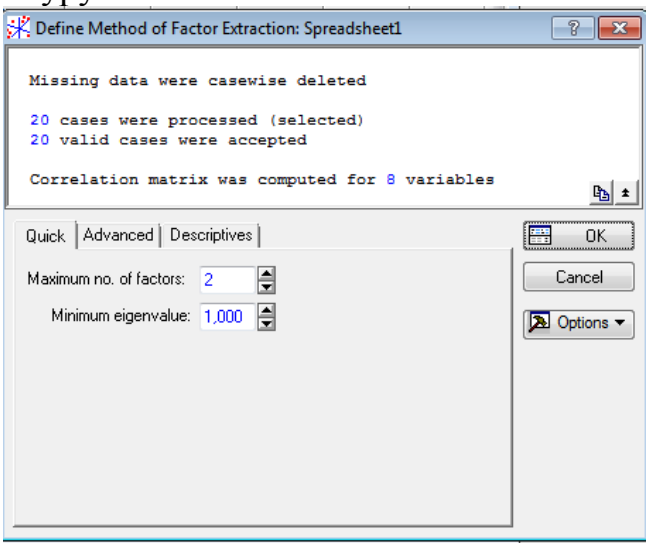
В модуле возможны следующие типы исходных данных:
Correlation Matrix (Корреляционная матрица);

Raw Data (Исходные данные).

· Выберите, например, **Raw Data**. Это обычный файл данных, где по строкам записаны значения переменных.

· В строке **Missing Data (Пропущенные данные)** задайте способ обработки пропущенных значений.

Casewise – способ исключения пропущенных случаев (состоит в том, что в электронной таблице, содержащей данные, игнорируются все строки (случаи), в которых имеется хотя бы одно пропущенное значение. Это относится ко всем переменным. В таблице остаются только случаи, в которых нет ни одного пропуска). Щелкните на кнопку **OK**, перед вами появиться окно **Define Method of Factor Extraction (Определить метод выделения факторов)**. Данное окно имеет следующую структуру:



Maximum no.of factors – Максимальное число факторов;

Minimum eigenvalue – Минимальное собственное значение.

Эти поля определяют число факторов, которые будут выделены системой. Собственные значения, меньше указанного в поле игнорируются.

Инициировав кнопку **Review corrs/means/SD (Просмотреть корреляции/средние/стандартные отклонения)**, вы откроете окно **Просмотреть описательные статистики.**

С помощью кнопки **Perform multiple regression (Выполнить множественную регрессию)** вы можете выполнить множественную регрессию, не выходя из модуля.

Группа опций, объединенных под заголовком **Extraction method** –метод выделения, позволяет выбрать метод обработки.

Как говорилось в математическом анализе, в зависимости от критерия оптимальности возможен анализ либо методом **Principal components** – методом главных компонент, либо одним из методов, объединенных в группу **Principal factor analysis – анализ главных (общих) факторов.**

Система предлагает следующие методы в группе **Principal factor analysis– анализ главных факторов**:

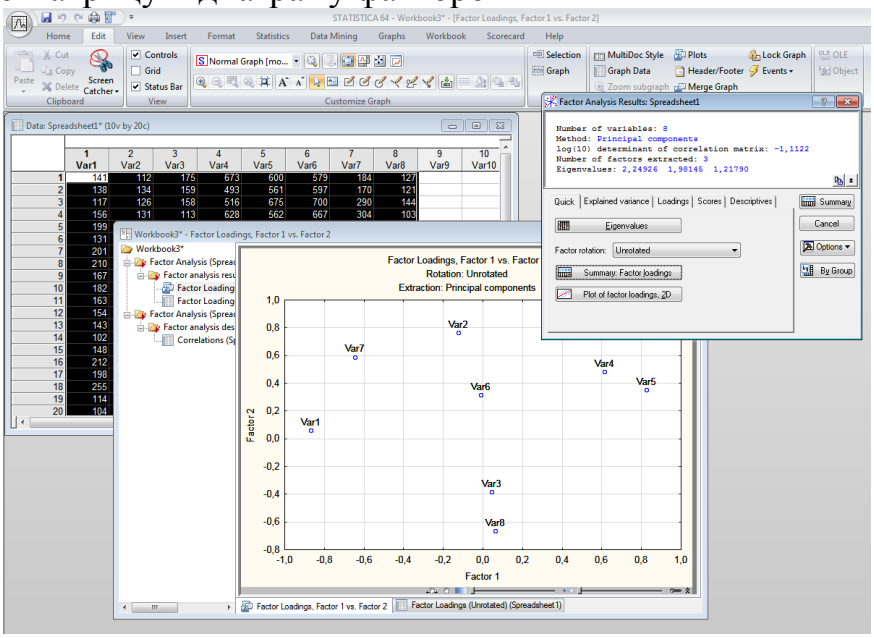
Communalities = multiple R2** (общности равны квадрату коэффициента множественной корреляции);

Iterated communalities (MINRES) – метод итеративных общностей (минимальных остатков);

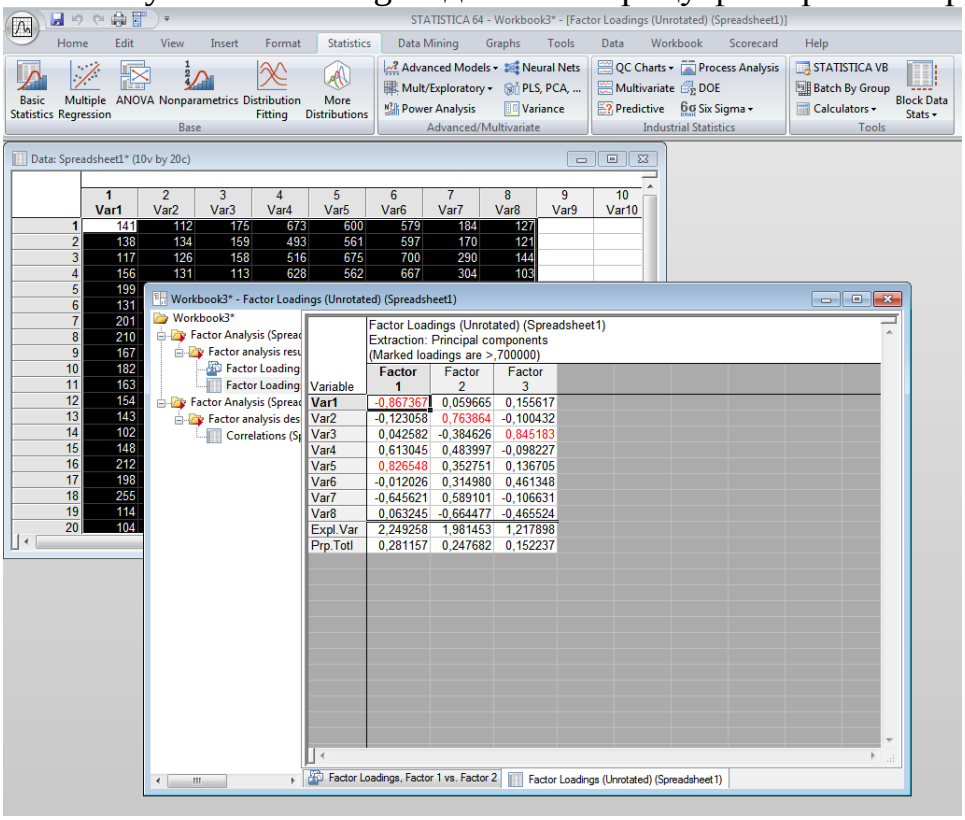
Centroid method – центроидный метод;

Principal axis method – метод главных осей.

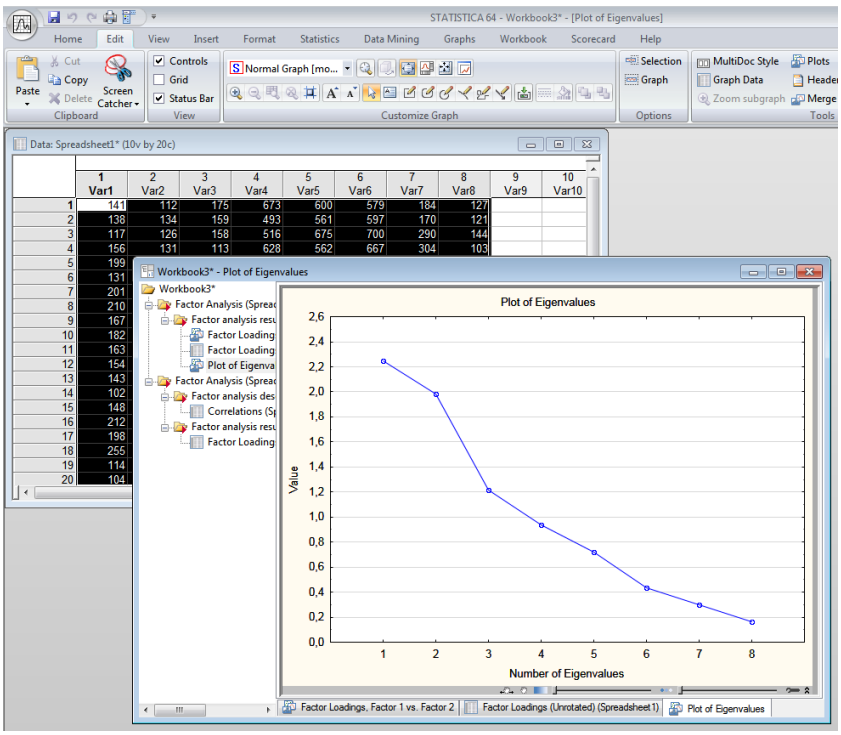
4. Выбрав **Principal components** открываем окно анализа и троим корреляционную матрицу и диаграму факторов



Выбрав Summary: factor loading выделяем матрицу факторных нагрузок



5. Используя критерий каменистой осыпи Кеттлера определяем оптимальное количество факторов:



Вариант 1

n	X1	X2	X3	X4	X5	X6	X7	X8
1	165	119	303	342	136	334	184	128
2	203	222	175	182	245	126	170	167
3	110	132	159	242	305	142	290	286
4	108	202	158	179	252	101	304	214
5	235	104	277	325	154	106	221	211
6	110	311	115	304	107	211	111	133
7	212	244	346	307	162	136	195	161
8	198	315	318	276	282	146	221	188
9	255	132	198	234	331	330	112	103
10	228	113	135	284	347	296	153	243
11	281	249	213	257	339	258	208	242
12	339	275	182	264	125	293	213	293
13	311	256	223	166	194	243	210	310
14	349	315	235	183	101	283	152	248
15	129	280	170	174	145	140	246	336
16	170	227	131	316	326	214	324	208
17	185	191	264	193	336	120	248	190
18	209	267	176	284	326	114	342	251
19	258	136	135	119	212	200	111	102
20	276	230	302	330	254	122	233	210

Вариант 2

n	X1	X2	X3	X4	X5	X6	X7	X8
1	267	741	456	548	323	671	237	514
2	639	444	660	311	797	308	701	536
3	707	521	521	629	798	266	272	710
4	466	563	755	294	526	606	386	259
5	665	664	645	378	627	418	671	602
6	341	438	384	785	515	478	581	453
7	716	512	201	496	654	570	205	598
8	478	496	454	276	428	501	250	342
9	431	624	303	663	582	575	601	632
10	722	474	352	628	630	353	454	661
11	246	756	278	512	253	478	623	536
12	622	388	417	241	458	387	226	432
13	344	233	745	607	321	286	342	413
14	710	316	215	435	700	209	244	335
15	723	724	687	416	527	346	612	416
16	772	766	417	421	669	249	510	688
17	677	620	530	485	374	516	750	520
18	526	652	468	674	678	340	421	618
19	668	205	411	416	401	545	403	299
20	296	673	469	315	312	690	645	239

Вариант 3

n	X1	X2	X3	X4	X5	X6	X7	X8
1	120	122	52	189	166	91	105	194
2	55	56	53	109	108	23	111	116
3	95	23	108	151	37	51	140	139
4	151	90	66	77	196	111	172	55
5	33	27	148	97	59	43	140	90
6	75	117	137	34	99	50	172	172
7	84	185	135	140	62	96	50	130
8	189	93	157	133	63	56	30	88
9	197	71	28	52	172	58	172	166
10	126	181	110	63	104	133	123	23
11	143	152	160	97	118	70	111	40
12	70	147	100	62	31	184	53	190
13	78	105	184	65	156	138	54	184
14	62	190	190	129	190	121	105	187
15	22	183	172	102	188	68	78	109
16	20	154	62	137	44	164	194	187
17	67	75	100	106	146	113	165	140
18	124	182	126	126	179	47	63	154

19	36	92	51	197	193	193	50	86
20	117	108	97	53	165	33	30	133

Вариант 4

n	X1	X2	X3	X4	X5	X6	X7	X8
1	156	111	117	79	44	96	144	127
2	122	172	178	60	78	52	111	121
3	181	187	148	77	43	46	117	144
4	110	195	179	98	72	63	103	103
5	110	117	187	66	41	84	143	118
6	165	136	160	83	67	87	147	115
7	154	189	147	55	83	54	121	122
8	181	106	149	77	40	81	138	139
9	127	154	199	52	62	54	109	137
10	104	157	191	71	60	56	147	133
11	163	108	192	97	64	46	142	145
12	141	178	198	74	93	98	135	144
13	143	135	192	59	42	85	130	105
14	143	116	100	94	45	50	131	118
15	153	169	163	91	42	61	140	108
16	102	137	189	93	48	61	120	108
17	196	135	135	100	43	81	127	134
18	136	115	128	75	49	90	147	109
19	113	126	108	85	73	83	144	140
20	173	171	163	61	50	72	129	128

Вариант 5

n	X1	X2	X3	X4	X5	X6	X7	X8
1	117	122	62	55	418	271	252	346
2	140	133	69	42	448	267	480	335
3	144	122	45	70	471	212	247	327
4	142	137	35	61	471	357	279	335
5	111	103	64	43	474	380	478	328
6	128	106	33	36	467	244	396	345
7	128	120	37	39	412	474	387	322
8	115	126	53	59	466	499	373	320
9	101	130	64	37	408	266	293	339
10	111	148	51	54	420	213	308	393
11	122	147	58	57	471	325	316	312
12	147	104	64	34	401	420	231	325
13	105	126	45	65	466	421	342	394
14	137	123	49	58	439	454	241	339
15	117	129	61	41	472	384	210	352

16	129	144	62	32	489	410	297	307
17	119	112	30	33	422	365	442	394
18	111	150	40	42	417	298	493	340
19	148	134	61	32	456	428	369	359
20	131	112	43	38	400	496	491	337

Вариант 6

n	X1	X2	X3	X4	X5	X6	X7	X8
1	331	345	376	44	26	136	139	116
2	340	387	360	57	34	114	148	133
3	363	345	396	40	29	143	126	100
4	307	371	360	36	51	140	111	118
5	366	349	395	59	20	139	112	107
6	306	376	373	46	26	131	122	128
7	371	333	342	58	27	144	113	111
8	384	352	321	39	52	142	135	113
9	371	375	370	54	21	134	131	144
10	390	391	348	51	36	123	130	110
11	333	336	309	20	58	125	123	102
12	334	320	351	22	26	120	124	139
13	317	307	309	39	34	115	141	134
14	347	383	322	46	37	120	113	115
15	323	345	330	26	23	149	142	119
16	378	300	360	56	31	131	149	148
17	368	343	397	29	48	111	112	116
18	321	376	322	26	51	131	147	101
19	319	371	380	43	56	147	117	150
20	330	393	387	60	25	103	135	108

Вариант 7

n	X1	X2	X3	X4	X5	X6	X7	X8
1	148	112	131	673	600	579	51	52
2	118	134	117	493	561	597	53	44
3	116	126	138	516	675	700	54	56
4	110	131	113	628	562	667	40	53
5	139	133	108	705	565	711	45	56
6	135	108	126	404	768	595	49	44
7	129	111	141	360	568	463	48	59
8	105	119	123	370	669	578	57	44
9	105	118	112	604	506	383	48	53
10	115	104	105	673	788	500	57	50
11	148	111	124	389	311	531	53	52
12	126	126	100	397	523	319	42	48

13	143	129	125	619	795	381	57	51
14	102	125	132	505	579	493	47	43
15	148	121	133	663	693	401	53	57
16	132	110	144	391	318	367	59	55
17	127	121	145	408	439	484	60	45
18	139	139	120	393	412	735	40	47
19	114	113	133	515	508	545	47	60
20	104	120	146	592	594	545	60	52

Вариант 8

n	X1	X2	X3	X4	X5	X6	X7	X8
1	44	42	166	181	330	392	39	55
2	44	42	243	106	430	201	51	43
3	47	46	203	255	287	456	58	51
4	45	54	113	171	332	225	29	29
5	58	57	117	225	397	288	16	24
6	56	50	238	107	262	406	27	20
7	44	56	150	151	273	444	26	60
8	49	49	237	235	203	326	19	52
9	45	53	201	200	395	219	21	42
10	50	60	249	150	243	448	37	15
11	43	54	149	227	224	263	39	30
12	46	60	105	232	288	271	33	41
13	49	51	194	139	296	285	16	53
14	41	59	198	227	322	325	55	48
15	56	59	106	104	327	253	29	53
16	52	56	125	113	333	253	20	49
17	54	60	142	231	372	303	48	57
18	55	41	247	220	317	456	55	55
19	50	52	107	158	269	261	20	52
20	59	49	160	202	301	426	24	37

Вариант 9

n	X1	X2	X3	X4	X5	X6	X7	X8
1	41	40	166	181	330	1872	1772	1196
2	18	25	243	106	430	1642	1377	1777
3	58	55	203	255	287	1798	1465	1586
4	55	30	113	171	332	1419	1141	1388
5	38	50	117	225	397	1677	1404	1904
6	17	49	238	107	262	1494	1562	1015
7	50	52	150	151	273	1661	1269	1631
8	53	41	237	235	203	1023	1791	1021
9	50	56	201	200	395	1140	1637	1233

10	18	42	249	150	243	1724	1281	1419
11	35	47	149	227	224	1381	1471	1471
12	45	39	105	232	288	1552	1889	1168
13	22	49	194	139	296	1875	1386	1681
14	43	28	198	227	322	1643	1711	1644
15	26	37	106	104	327	1932	1820	1364
16	54	46	125	113	333	1223	1315	1105
17	56	17	142	231	372	1832	1967	1285
18	22	43	247	220	317	1663	1542	1344
19	41	19	107	158	269	1686	1713	1068
20	45	43	160	202	301	1416	1465	1524

Вариант 10

n	X1	X2	X3	X4	X5	X6	X7	X8
1	1734	1041	1725	148	172	152	19	14
2	1347	1497	1375	137	199	124	38	17
3	1139	1004	1379	193	186	162	29	30
4	1520	1268	1110	142	104	176	24	36
5	1042	1223	1174	148	148	119	21	17
6	1054	1080	1911	177	185	180	29	33
7	1458	1905	1188	185	199	140	14	22
8	1307	1020	1243	110	138	129	15	12
9	1277	1213	1349	112	106	126	13	22
10	1423	1318	1102	129	158	121	19	32
11	1365	1166	1198	143	189	135	24	27
12	1679	1358	1882	150	191	108	37	19
13	1361	1079	1031	183	109	139	27	28
14	1645	1525	1834	179	174	185	20	13
15	1211	1204	1257	187	157	125	26	36
16	1291	1226	1964	129	110	192	21	14
17	1674	1620	1506	108	111	191	34	10
18	1314	1064	1075	195	163	134	40	19
19	1424	1234	1707	108	135	127	37	38
20	1666	1909	1568	112	196	118	29	16

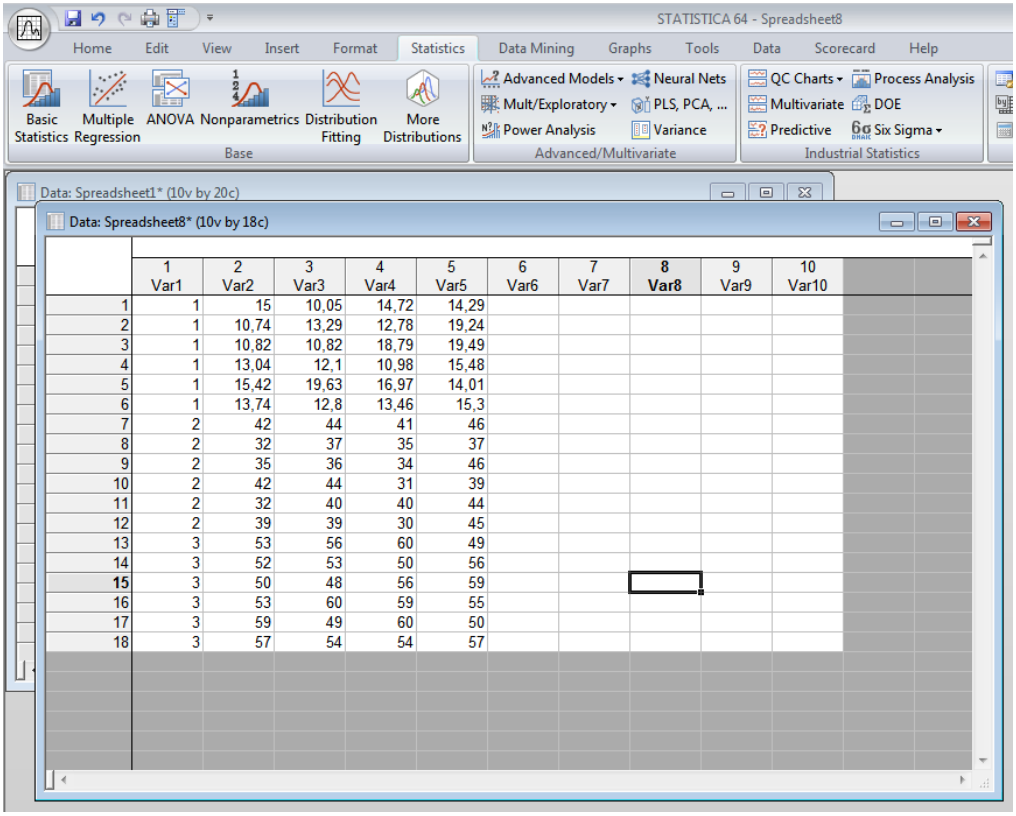
Практическое занятие № 4 Построение дискриминантных моделей

По предложенному варианту построить дискриминантную разделительную функцию используя ППП Statistica или SPSS. Классифицировать по полученной дискриминационной модели неклассифицированные объекты.

k - классы, n - объекты классификации, X1, X2, X3, X4 - классификационные функции.

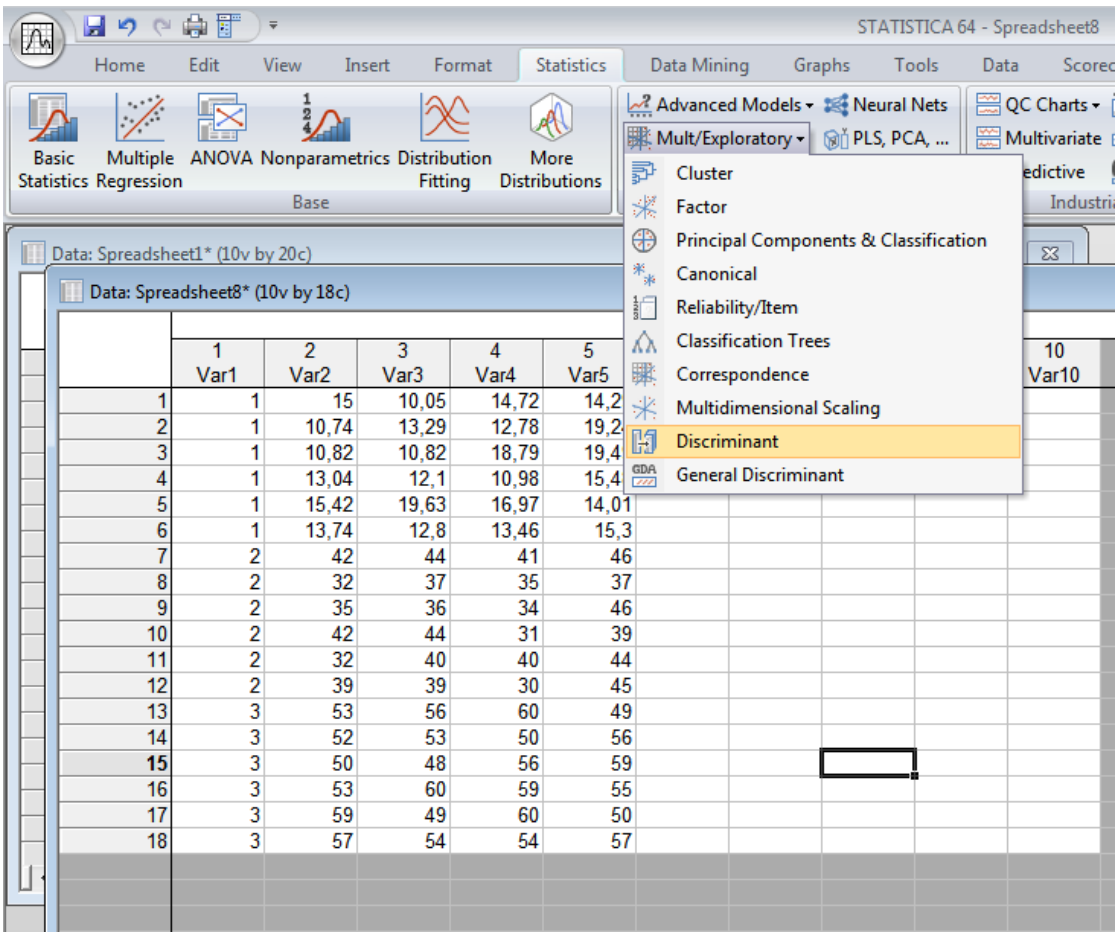
n	k	x1	x2	x3	x4
n1	1	15,00	10,05	14,72	14,29
n2	1	10,74	13,29	12,78	19,24
n3	1	10,82	10,82	18,79	19,49
n4	1	13,04	12,10	10,98	15,48
n5	1	15,42	19,63	16,97	14,01
n6	1	13,74	12,80	13,46	15,30
n7	2	42	44	41	46
n8	2	32	37	35	37
n9	2	35	36	34	46
n10	2	42	44	31	39
n11	2	32	40	40	44
n12	2	39	39	30	45
n13	3	53	56	60	49
n14	3	52	53	50	56
n15	3	50	48	56	59
n16	3	53	60	59	55
n17	3	59	49	60	50
n18	3	57	54	54	57
неклассифицированные объекты					
n19		23	51	53	56
n20		48	26	43	14
n21		38	31	22	11
n22		54	26	41	21
n23		23	22	34	37
n24		53	57	20	47

1. Запускаем программу Statistica и загружаем исходные данные:



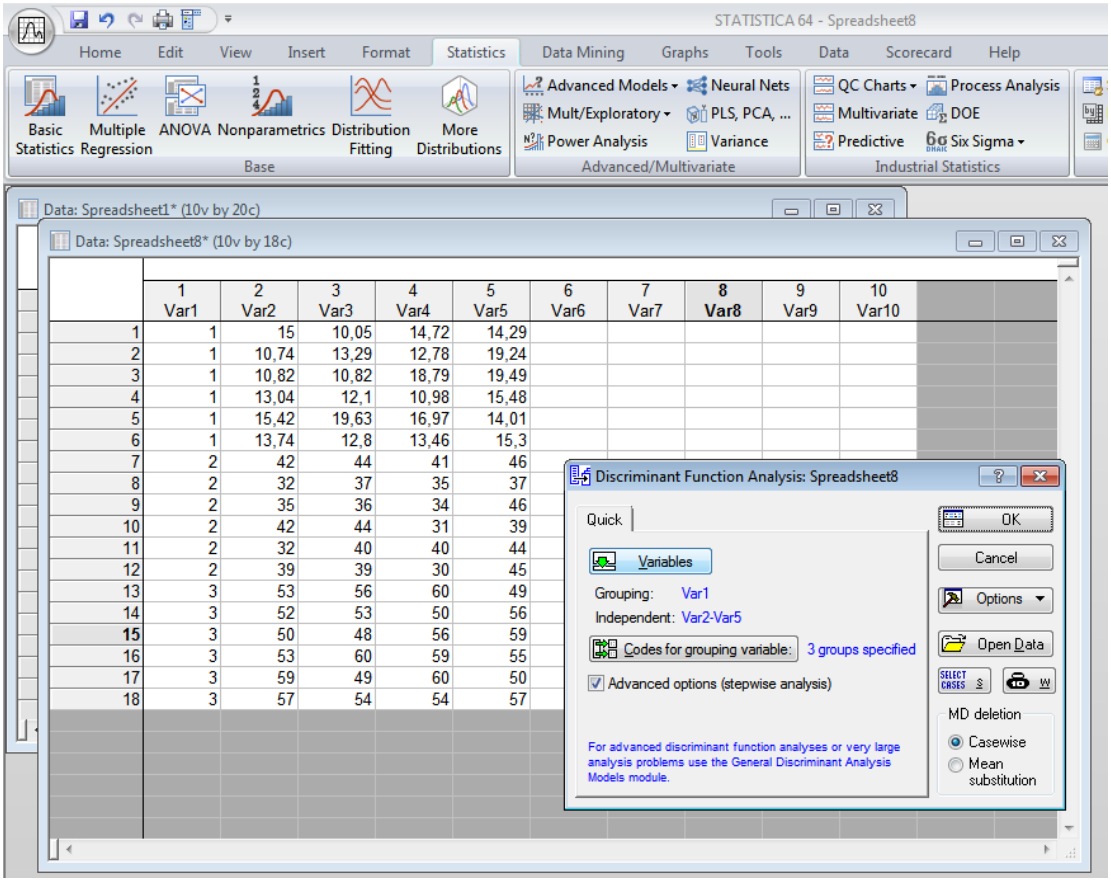
	1 Var1	2 Var2	3 Var3	4 Var4	5 Var5	6 Var6	7 Var7	8 Var8	9 Var9	10 Var10
1	1	15	10,05	14,72	14,29					
2	1	10,74	13,29	12,78	19,24					
3	1	10,82	10,82	18,79	19,49					
4	1	13,04	12,1	10,98	15,48					
5	1	15,42	19,63	16,97	14,01					
6	1	13,74	12,8	13,46	15,3					
7	2	42	44	41	46					
8	2	32	37	35	37					
9	2	35	36	34	46					
10	2	42	44	31	39					
11	2	32	40	40	44					
12	2	39	39	30	45					
13	3	53	56	60	49					
14	3	52	53	50	56					
15	3	50	48	56	59					
16	3	53	60	59	55					
17	3	59	49	60	50					
18	3	57	54	54	57					

2. В надстройке Statistic выбираем анализ Mult/Exporatory - Discriminant



	1 Var1	2 Var2	3 Var3	4 Var4	5 Var5	10 Var10
1	1	15	10,05	14,72	14,2	
2	1	10,74	13,29	12,78	19,2	
3	1	10,82	10,82	18,79	19,4	
4	1	13,04	12,1	10,98	15,4	
5	1	15,42	19,63	16,97	14,01	
6	1	13,74	12,8	13,46	15,3	
7	2	42	44	41	46	
8	2	32	37	35	37	
9	2	35	36	34	46	
10	2	42	44	31	39	
11	2	32	40	40	44	
12	2	39	39	30	45	
13	3	53	56	60	49	
14	3	52	53	50	56	
15	3	50	48	56	59	
16	3	53	60	59	55	
17	3	59	49	60	50	
18	3	57	54	54	57	

Далее выберем *Общий дискриминантный анализ* и диалоговый способ задания параметров анализа.

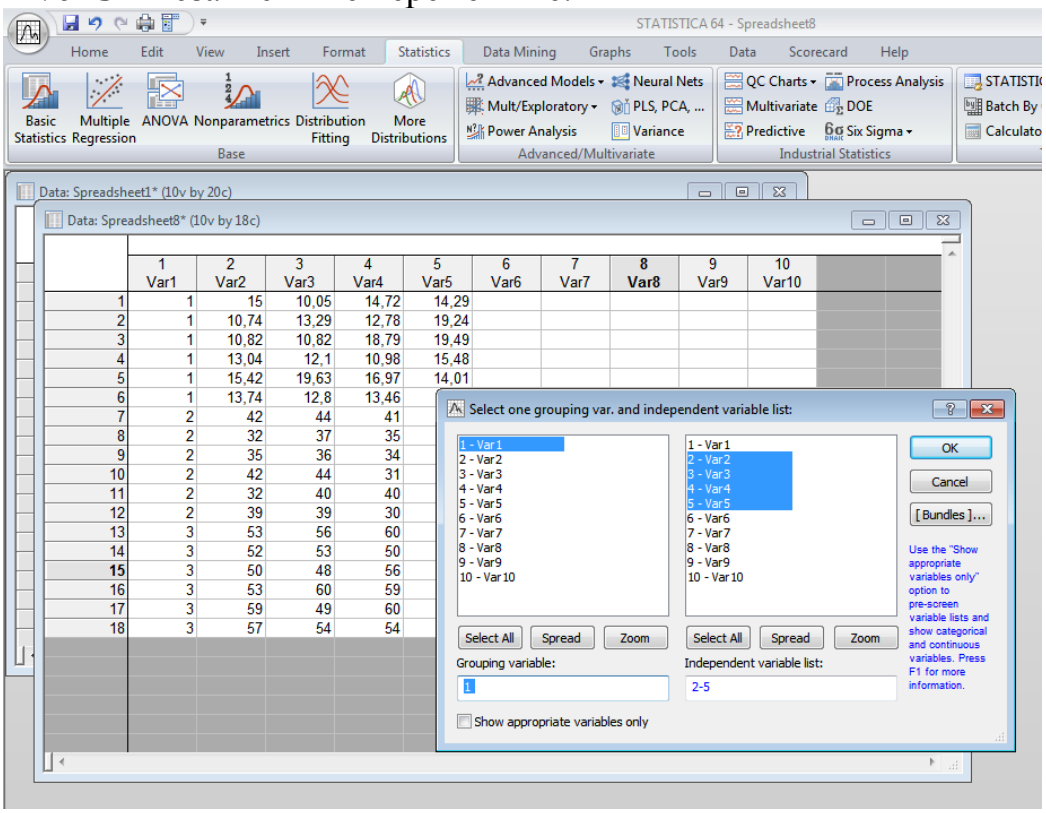


3. Теперь зададим параметры анализа.

Нажав кнопку Переменные, зададим переменные анализа.

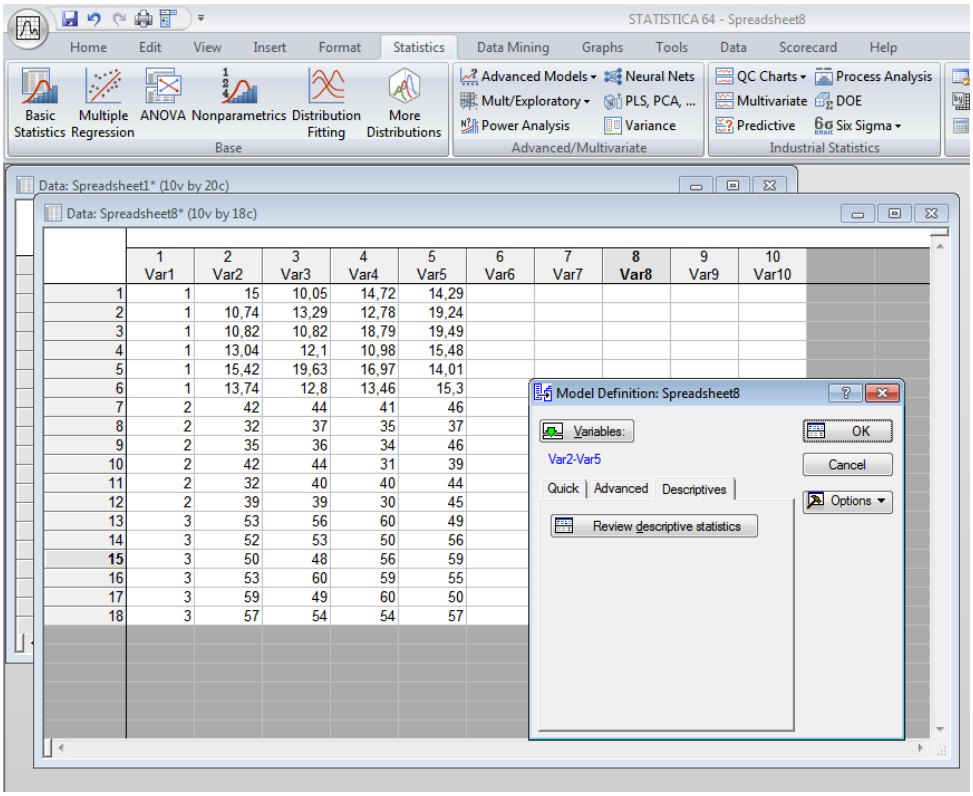
Var1 – группирующая переменная;

Var2 - Var 5 – независимые переменные.

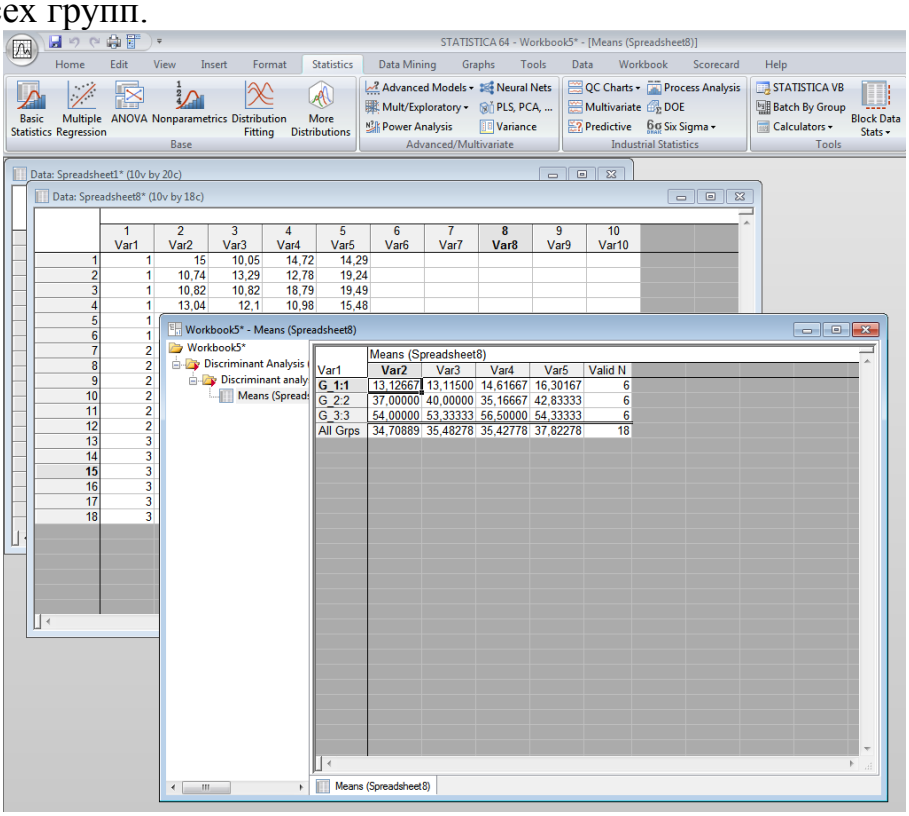


4. Просмотр описательных статистик. Теперь, для того, чтобы начать анализ, нажмите на кнопкуOK. Откроется диалоговое окно Определение

модели, которое позволит задать параметры дискриминантного анализа и просмотреть описательные статистики.



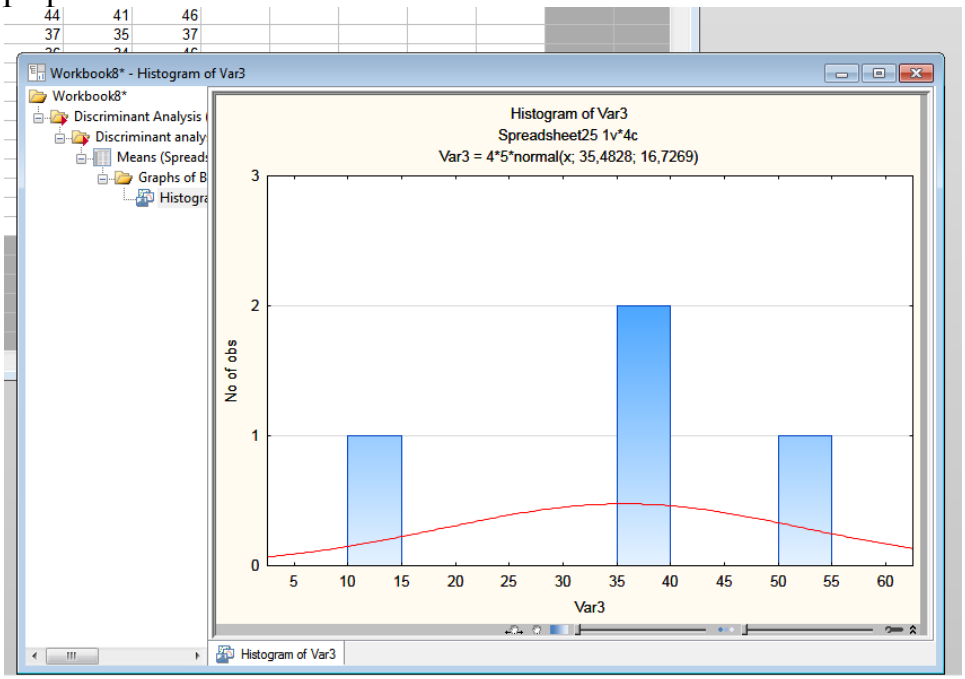
5. Сначала взгляните на средние. Во вкладке Быстрый нажмите на кнопку Средние и число наблюдений, после чего должна появиться таблица со средними и действующим значением n для каждой совокупности и для комбинации всех групп.



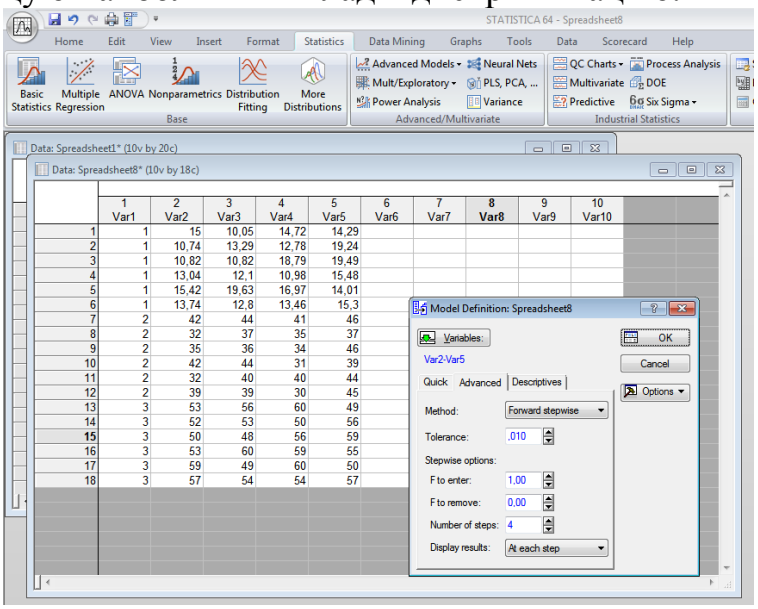
6. Получение гистограммы из таблицы результатов. Для получения гистограммы частот распределения для какой-либо переменной, выделите

сначала нужный столбец в таблице. Вы можете получить или гистограмму для всех групп или только для выбранных групп.

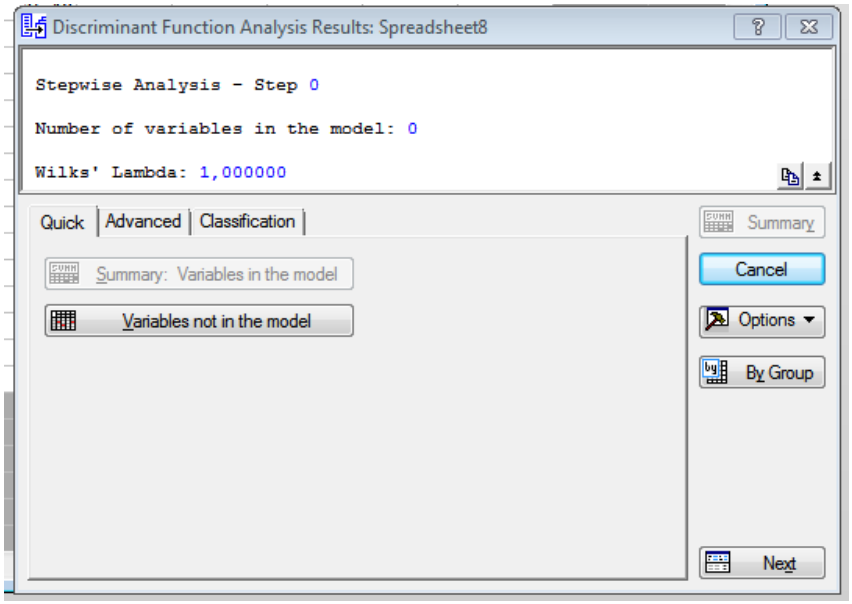
Например, для получения гистограммы переменной Sepalwid только для типа Versicoove, передвиньте курсор на пересечение второй строки и второго столбца таблицы. Затем нажмите на правую кнопку мыши, чтобы открыть контекстное меню и выбрать команду Графики исходных данных. Теперь выберите команду Гистограмма - Нормальная подгонка, и вы получите следующий график.



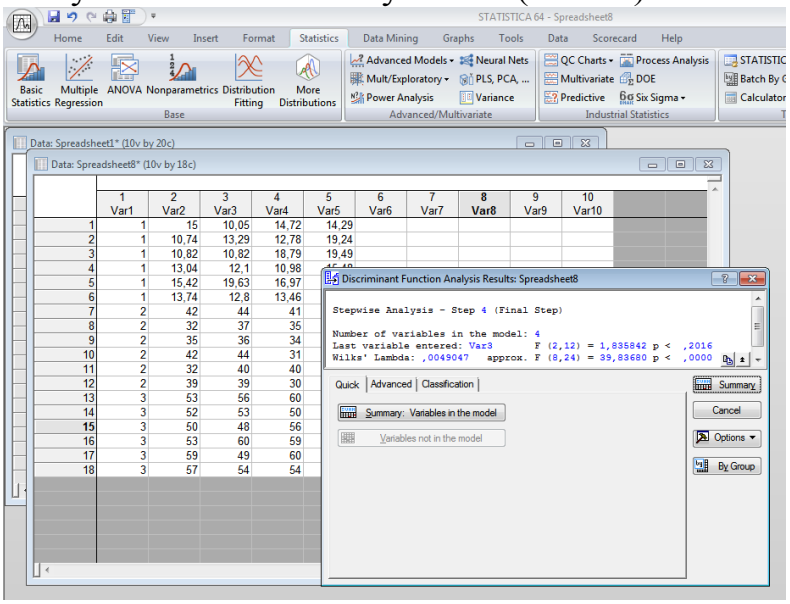
7. Во вкладке Дополнительно, в списке Метод установите значение Пошаговый с включением. При такой установке программа будет вводить переменные в модель постепенно, одну за другой, каждый раз выбирая переменную, вносящую наибольший вклад в дискриминацию.



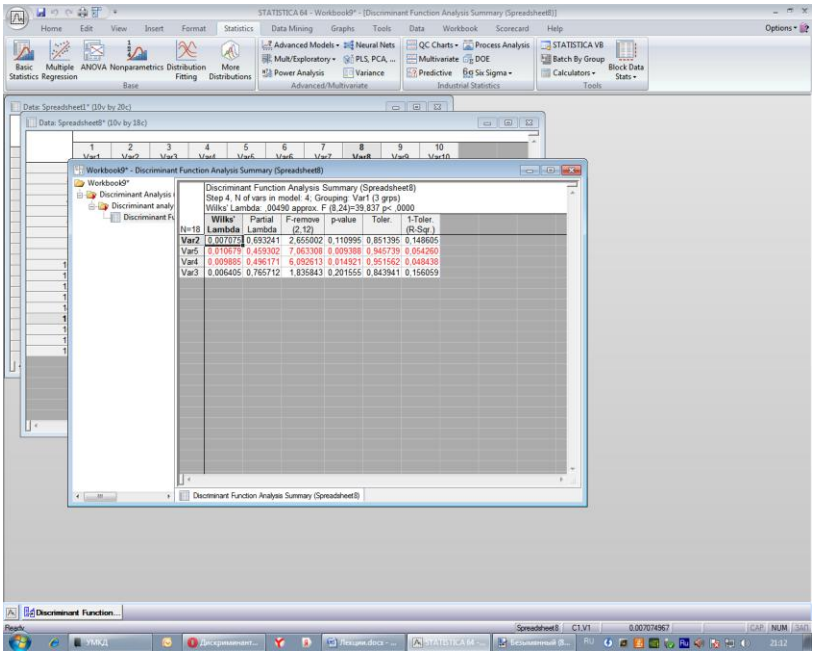
Результаты на шаге 0. Сначала отображаются Результаты дискриминантного анализа на нулевом шаге. Слова Шаг 0 означают, что еще ни одной переменной в модель не было включено.



Нажимаем кнопку next по количеству шагов (4 шага)



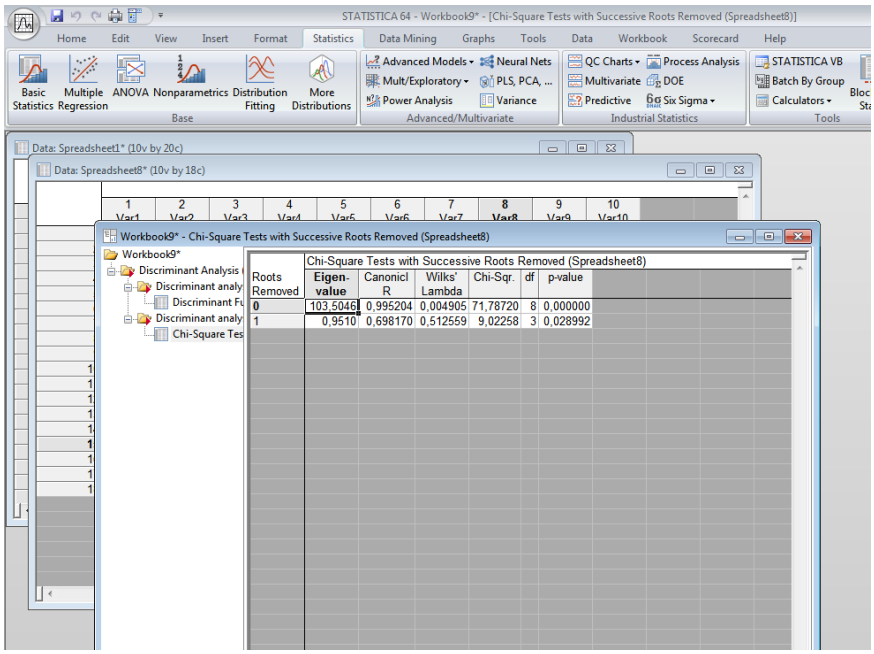
Теперь нажмем на кнопку Переменные в модели для обзора независимых вкладов каждой переменной в общую дискриминацию между типами ирисов.



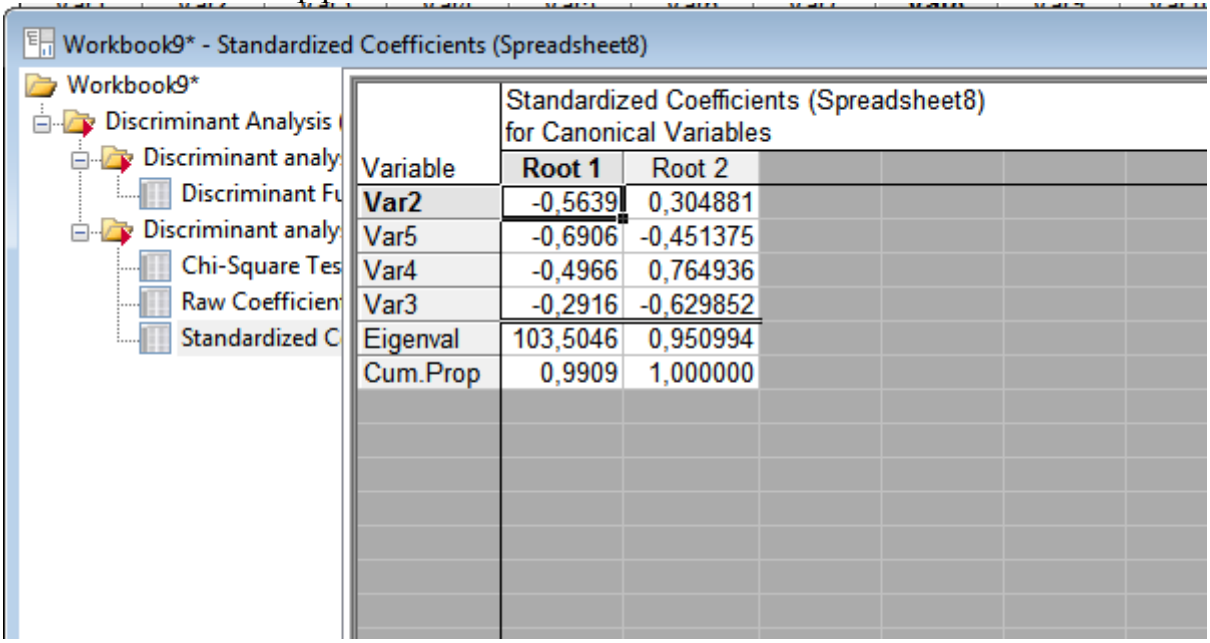
Для получения дальнейших результатов о природе дискриминации следует провести канонический анализ. Стартовую панель Канонического анализа можно вызвать соответствующей кнопкой во вкладке Дополнительно диалогового окна Результаты дискриминантного анализа.

8. Канонический анализ. Чтобы увидеть, как четыре переменные разделяют различные совокупности (типы ирисов), вычислим действительную дискриминантную функцию. Нажмите на кнопку Канонический анализ для выполнения канонического анализа и откройте диалоговое окно Канонический анализ.

Значимость корней. Сначала определим, являются ли обе дискриминантные функции (корни) статистически значимыми. Нажмите на кнопку Критерий Хи-квадрат последовательных корней и увидите следующую таблицу:



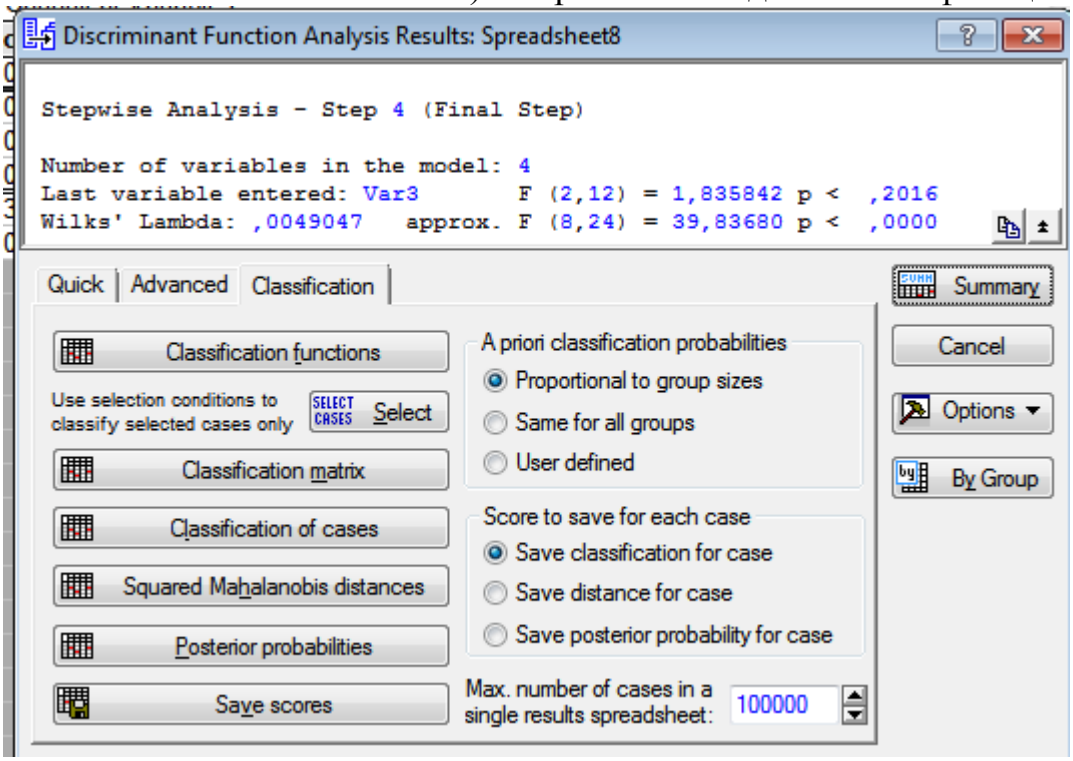
9. Коэффициенты дискриминантной функции. Нажмите на кнопку Коэффициенты для канонических переменных в диалоговом окне Канонический анализ. Будут получены две таблицы, одна для Исходных коэффициентов и другая для Стандартизованных коэффициентов. Взглянем теперь на Исходные коэффициенты.

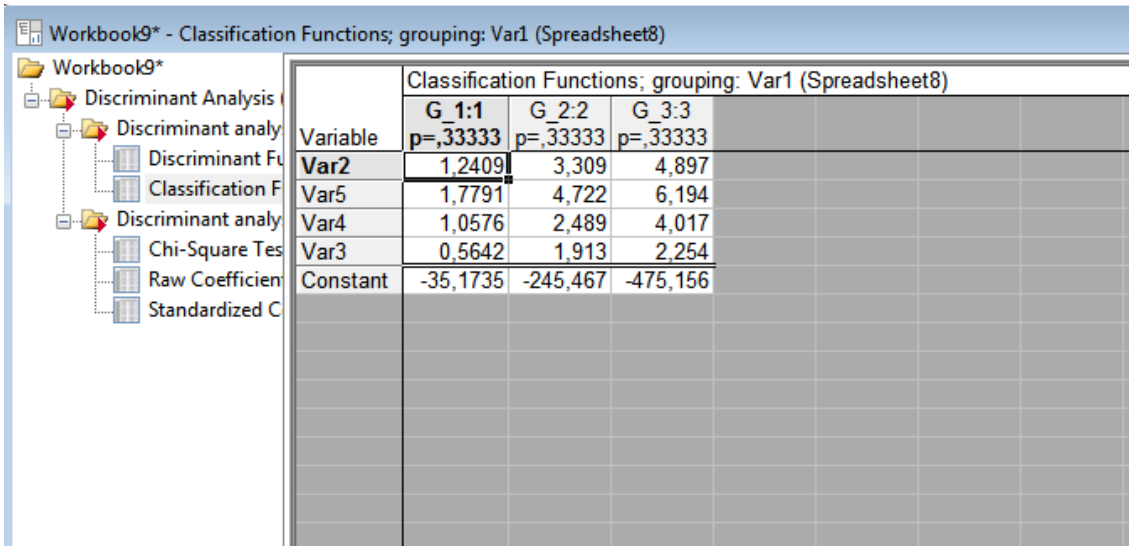


Standardized Coefficients (Spreadsheet8) for Canonical Variables

Variable	Root 1	Root 2
Var2	-0,5639	0,304881
Var5	-0,6906	-0,451375
Var4	-0,4966	0,764936
Var3	-0,2916	-0,629852
Eigenval	103,5046	0,950994
Cum.Prop	0,9909	1,000000

10. Для решения задачи классификации вернемся в диалоговое окно Результаты анализа дискриминантных функций (нажмите на кнопку Отмена в диалоговом окне Канонический анализ) и вернитесь к задаче классификации.





Classification Functions; grouping: Var1 (Spreadsheet8)

Variable	G_1:1 p=,33333	G_2:2 p=,33333	G_3:3 p=,33333
Var2	1,2409	3,309	4,897
Var5	1,7791	4,722	6,194
Var4	1,0576	2,489	4,017
Var3	0,5642	1,913	2,254
Constant	-35,1735	-245,467	-475,156

С помощью этих функций можно будет в дальнейшем классифицировать новые случаи. Новые случаи будут относиться к тому классу, для которого классифицированное значение будет максимальное. Выбор метода окончательной классификации зависит от количества новых объектов, подлежащих классификации. Если количество новых случав невелико, можно применить метод, основанный на статистических критериях. Если же количество новых случаев велико, то рациональнее по обучающим выборкам получить классификационные функции и затем, настроить формулы и провести окончательную классификацию.

Вариант 1

n	k	x1	x2	x3	x4
n1	1	24	10	17	33
n2	1	34	27	19	12
n3	1	25	33	28	12
n4	1	37	19	19	32
n5	1	29	26	33	18
n6	1	18	25	20	32
n7	2	42	44	41	46
n8	2	32	37	35	37
n9	2	35	36	34	46
n10	2	42	44	31	39
n11	2	32	40	40	44
n12	2	39	39	30	45
n13	3	53	56	60	49
n14	3	52	53	50	56
n15	3	50	48	56	59
n16	3	53	60	59	55
n17	3	59	49	60	50

n18	3	57	54	54	57
неклассифицированные объекты					
n19		23	51	53	56
n20		48	26	43	14
n21		38	31	22	11
n22		54	26	41	21
n23		23	22	34	37
n24		53	57	20	47

Вариант 2

n	k	x1	x2	x3	x4
n1	1	15,00	10,05	14,72	14,29
n2	1	10,74	13,29	12,78	19,24
n3	1	10,82	10,82	18,79	19,49
n4	1	13,04	12,10	10,98	15,48
n5	1	15,42	19,63	16,97	14,01
n6	1	13,74	12,80	13,46	15,30
n7	2	20,83	21,59	21,39	23,82
n8	2	24,15	23,44	22,85	20,72
n9	2	22,56	22,06	22,00	23,97
n10	2	23,83	20,87	20,50	20,04
n11	2	21,68	21,89	20,52	20,47
n12	2	24,81	21,53	22,68	21,21
n13	3	28,89	25,59	24,60	28,56
n14	3	27,66	24,52	27,13	25,70
n15	3	26,00	24,32	24,79	24,84
n16	3	24,06	26,10	24,66	24,76
n17	3	28,62	26,98	27,69	24,33
n18	3	28,64	28,40	25,82	26,21
неклассифицированные объекты					
n19		17,50	10,09	29,92	11,68
n20		10,53	28,06	23,86	29,55
n21		28,38	28,61	14,49	17,08
n22		18,62	26,03	17,96	18,88
n23		18,33	26,27	14,96	27,31
n24		17,17	10,13	23,60	16,31

Вариант 3

n	k	x1	x2	x3	x4
n1	1	68,47	50,91	41,70	65,55
n2	1	48,70	60,10	53,41	65,65
n3	1	53,37	59,13	53,35	56,25
n4	1	45,63	54,03	52,00	55,28

n5	1	48,66	59,38	68,11	43,88
n6	1	51,86	53,61	41,70	51,17
n7	2	61,53	72,54	60,51	74,58
n8	2	80,40	77,85	88,79	67,18
n9	2	80,16	86,42	67,91	60,07
n10	2	86,46	65,36	69,46	84,70
n11	2	61,44	63,69	64,92	66,12
n12	2	66,39	73,35	76,23	80,19
n13	3	96,61	89,86	87,76	92,33
n14	3	87,59	86,01	97,40	98,18
n15	3	88,88	91,41	96,09	94,37
n16	3	87,69	90,34	90,74	98,52
n17	3	88,69	98,77	86,07	96,03
n18	3	99,51	89,60	93,36	90,66
неклассифицированные объекты					
n19		72,00	80,75	99,53	77,96
n20		98,69	77,43	81,37	90,42
n21		86,74	99,06	53,20	82,70
n22		92,78	53,67	87,04	60,40
n23		65,13	51,01	79,69	45,20
n24		52,47	53,89	85,96	72,73

Вариант 4

n	k	x1	x2	x3	x4
n1	1	28,52	23,49	24,20	24,06
n2	1	11,78	16,80	17,02	21,14
n3	1	12,55	22,55	23,29	28,75
n4	1	19,78	22,92	19,32	19,20
n5	1	22,65	22,03	27,73	29,90
n6	1	23,28	20,42	16,37	18,87
n7	2	32,13	36,47	34,77	32,63
n8	2	34,54	30,98	39,24	39,34
n9	2	37,48	38,56	34,43	38,56
n10	2	29,80	30,67	33,97	31,31
n11	2	32,69	32,91	35,41	35,78
n12	2	36,98	38,74	34,61	33,79
n13	3	44,56	39,45	39,65	43,77
n14	3	40,24	42,46	39,59	43,74
n15	3	44,91	44,87	42,64	41,09
n16	3	41,17	43,42	41,95	43,36

n17	3	42,61	41,90	39,53	43,30
n18	3	41,53	40,65	39,55	39,81
неклассифицированные объекты					
n19		35,18	20,72	35,91	42,90
n20		12,77	17,60	21,57	21,52
n21		11,82	10,67	21,69	23,00
n22		42,08	10,04	10,54	23,51
n23		40,77	29,25	34,28	26,54
n24		43,92	13,21	11,71	17,61

Вариант 5

n	k	x1	x2	x3	x4
n1	1	43,91	43,39	41,05	40,01
n2	1	44,04	41,50	41,13	41,58
n3	1	43,65	40,49	43,34	44,52
n4	1	44,71	41,71	44,29	44,54
n5	1	41,13	42,93	44,62	40,25
n6	1	42,89	44,06	42,57	40,59
n7	2	46,93	45,08	46,94	46,13
n8	2	45,54	45,15	45,94	45,67
n9	2	47,87	45,61	45,29	45,31
n10	2	47,31	45,05	45,33	45,19
n11	2	47,37	46,47	46,04	46,96
n12	2	45,97	46,16	47,78	47,37
n13	3	51,24	51,74	49,21	51,69
n14	3	49,65	49,94	51,89	49,87
n15	3	50,97	50,49	51,21	50,68
n16	3	50,48	51,59	50,83	49,63
n17	3	49,51	50,95	51,02	49,24
n18	3	49,94	49,17	50,79	51,62
неклассифицированные объекты					
n19		45,71	41,38	43,58	48,00
n20		47,85	44,52	49,11	47,67
n21		44,58	41,37	42,40	41,32
n22		43,94	51,63	43,00	40,23
n23		48,41	48,67	50,30	41,91
n24		48,37	46,80	46,96	51,42

Вариант 6

n	k	x1	x2	x3	x4
n1	1	52,75	44,44	44,76	58,82
n2	1	58,16	46,03	43,31	42,95
n3	1	57,26	41,79	54,24	43,37

n4	1	52,74	55,74	59,27	43,70
n5	1	57,63	47,25	54,76	41,16
n6	1	51,09	56,28	58,75	58,24
n7	2	34,80	34,34	30,45	32,24
n8	2	38,39	34,68	38,61	32,45
n9	2	37,13	30,25	36,95	31,03
n10	2	34,20	39,52	30,98	33,82
n11	2	30,27	31,07	38,09	34,08
n12	2	33,99	38,38	34,55	37,99
n13	3	26,45	24,32	20,35	29,24
n14	3	28,83	24,70	27,95	26,17
n15	3	20,97	23,09	27,54	25,68
n16	3	20,07	25,58	22,64	26,65
n17	3	26,03	27,78	25,59	29,78
n18	3	28,74	29,14	25,65	29,45
неклассифицированные объекты					
n19		39,97	30,35	22,53	24,43
n20		28,41	22,54	26,48	31,98
n21		24,53	26,63	30,50	30,19
n22		21,85	30,71	34,38	37,40
n23		22,02	29,91	22,62	20,56
n24		32,83	24,75	31,75	25,68

Вариант 7

n	k	x1	x2	x3	x4
n1	1	99,99	88,08	96,40	90,59
n2	1	89,84	98,68	85,99	91,86
n3	1	93,14	95,82	88,53	97,24
n4	1	93,29	80,45	97,60	86,68
n5	1	98,56	93,45	83,44	80,83
n6	1	82,75	96,20	83,63	91,61
n7	2	66,55	77,09	65,66	76,61
n8	2	72,27	74,31	67,37	62,06
n9	2	72,19	61,70	75,80	68,78
n10	2	64,63	60,72	67,42	74,76
n11	2	63,43	76,37	77,66	72,50
n12	2	64,67	69,58	66,15	65,85
n13	3	52,38	58,79	45,39	55,93
n14	3	59,90	51,44	45,74	48,06
n15	3	59,39	58,75	46,06	41,70

n16	3	50,79	59,14	46,26	41,39
n17	3	57,70	40,25	58,24	57,05
n18	3	43,28	43,65	54,11	51,95
неклассифицированные объекты					
n19		94,22	59,80	71,80	53,33
n20		74,75	86,64	49,95	50,84
n21		82,59	73,40	84,77	63,35
n22		47,23	73,71	98,87	42,65
n23		84,48	96,61	45,92	95,28
n24		62,05	67,02	96,84	81,57

Вариант 8

n	k	x1	x2	x3	x4
n1	1	11,92	9,39	13,86	8,07
n2	1	12,68	11,79	6,67	7,33
n3	1	8,56	12,96	7,31	11,00
n4	1	6,54	14,61	7,79	6,67
n5	1	8,10	14,68	12,12	13,79
n6	1	5,36	10,20	13,99	11,87
n7	2	14,80	12,34	14,74	17,47
n8	2	19,41	18,54	15,37	10,00
n9	2	12,39	16,32	16,51	19,27
n10	2	18,24	12,80	19,87	15,94
n11	2	12,43	14,21	13,95	17,96
n12	2	15,86	18,64	17,10	17,56
n13	3	20,84	21,96	21,85	21,26
n14	3	20,36	20,48	20,64	21,26
n15	3	21,02	21,01	21,83	20,31
n16	3	21,23	21,01	20,15	21,91
n17	3	20,08	21,15	20,66	20,72
n18	3	21,47	21,27	20,18	20,11
неклассифицированные объекты					
n19		16,20	11,87	15,45	12,74
n20		19,91	14,25	12,21	12,37
n21		16,27	9,48	13,51	17,77
n22		16,38	7,51	17,14	5,04
n23		10,98	16,76	18,62	7,94
n24		10,47	8,27	15,50	9,12

Вариант 9

n	k	x1	x2	x3	x4
n1	1	52,75	44,44	44,76	58,82
n2	1	58,16	46,03	43,31	42,95

n3	1	57,26	41,79	54,24	43,37
n4	1	52,74	55,74	59,27	43,70
n5	1	57,63	47,25	54,76	41,16
n6	1	51,09	56,28	58,75	58,24
n7	2	34,80	34,34	30,45	32,24
n8	2	38,39	34,68	38,61	32,45
n9	2	37,13	30,25	36,95	31,03
n10	2	34,20	39,52	30,98	33,82
n11	2	30,27	31,07	38,09	34,08
n12	2	33,99	38,38	34,55	37,99
n13	3	26,45	24,32	20,35	29,24
n14	3	28,83	24,70	27,95	26,17
n15	3	20,97	23,09	27,54	25,68
n16	3	20,07	25,58	22,64	26,65
n17	3	26,03	27,78	25,59	29,78
n18	3	28,74	29,14	25,65	29,45
неклассифицированные объекты					
n19		39,97	30,35	22,53	24,43
n20		28,41	22,54	26,48	31,98
n21		24,53	26,63	30,50	30,19
n22		21,85	30,71	34,38	37,40
n23		22,02	29,91	22,62	20,56
n24		32,83	24,75	31,75	25,68

Вариант 10

n	k	x1	x2	x3	x4
n1	1	68,47	52,75	13,86	65,55
n2	1	48,70	58,16	6,67	65,65
n3	1	53,37	57,26	7,31	56,25
n4	1	45,63	52,74	7,79	55,28
n5	1	48,66	57,63	12,12	43,88
n6	1	51,86	51,09	13,99	51,17
n7	2	61,53	34,80	14,74	74,58
n8	2	80,40	38,39	15,37	67,18
n9	2	80,16	37,13	16,51	60,07
n10	2	86,46	34,20	19,87	84,70
n11	2	61,44	30,27	13,95	66,12
n12	2	66,39	33,99	17,10	80,19
n13	3	96,61	26,45	21,85	92,33
n14	3	87,59	28,83	20,64	98,18

n15	3	88,88	20,97	21,83	94,37
n16	3	87,69	20,07	20,15	98,52
n17	3	88,69	26,03	20,66	96,03
n18	3	99,51	28,74	20,18	90,66
неклассифицированные объекты					
n19		72,00	30,35	15,45	77,96
n20		98,69	22,54	12,21	90,42
n21		86,74	26,63	13,51	82,70
n22		92,78	30,71	17,14	60,40
n23		65,13	29,91	18,62	45,20
n24		52,47	24,75	15,50	72,73

Модуль 2. Основы моделирования систем

Тема 2.1 Модели прогнозирования

Лекция №5 Классические модели прогнозирования

План лекции

1. Методы прогнозирования и их классификация
2. Прогнозирование одномерных временных рядов
3. Эконометрические модели прогнозирования
4. Динамические модели прогнозирования

1. Методы прогнозирования и их классификация

Прогноз (от греческого prognosis – предвидение, предсказание) о состоянии какого-либо явления в будущем на основе специальных научных исследований. Прогноз неразрывно связан со временем и может быть определен различными методами – от простейших как, например, метод среднего уровня до динамических моделей прогнозирования.

Прогнозирование не может распространяться на те процессы и явления, результаты которых однозначны. Если же управление этими процессами и явлениями невозможно или возможно в весьма малом диапазоне прогнозирование необходимо.

Период упреждения прогноза – это отрезок времени от момента последних данных об изучаемом процессе или явлении до момента, к которому относится прогноз. По длительности периода упреждения выделяют следующие прогнозы:

краткосрочные – период упреждения до одного года;

среднесрочные – период упреждения от одного года до пяти лет;

долгосрочные – период упреждения свыше пяти лет.

Различают два вида прогноза – точечный и интервальный. Точечный прогноз – это прогноз в виде единственного значения прогнозируемой величины, а интервальный прогноз чаще всего определяется на основе расчета доверительного интервала с принятым уровнем доверительной вероятности на основе точечного прогноза по формуле: $Y_i = \widehat{Y} \pm t_\alpha S_q$, где Y_i – верхняя и нижняя граница интервального прогноза; $\widehat{Y}$ – точечный прогноз; t_α – значение статистики Стьюдента; S_q – ошибка прогноза.

Многие методы прогнозирования требуют большого количества исходных достоверных статистических данных и при их отсутствии они просто не могут быть использованы. Другие методы требуют минимального объема количественной информации. Поэтому, укрупнено, методы прогнозирования можно разбить на две группы:

качественные;

количественные.

Качественные методы основаны на использовании мнений специалистов в данной конкретной области. К ним можно отнести методы экспертных оценок. Суть этих методов следующая: в них используется такой показатель как ранг, характеризующий порядковое место оцениваемого параметра. Используется и такая процедура как ранжирование – т.е. установление относительной важности исследуемых параметров на основе их предпочтительности друг перед другом. Обычно наиболее предпочтительному параметру присваивается первый ранг, а наименее предпочтительному – последний ранг.

Степень согласованности оценок экспертов характеризуется коэффициентом конкордации (согласия) (W).

Коэффициент конкордации изменяется в пределах $0 \le W \le 1$. Согласованность считается удовлетворительной если, $W \ge 0{,}5$, если $W \ge 0{,}7$, то согласованность считается хорошей. При полном согласии экспертов $W = 1$.

Количественные методы прогнозирования основаны на обработке статистических данных и, в свою, очередь подразделяются на причинно-следственные (каузальные) методы и методы анализа одновременных временных рядов.

Классификация методов прогнозирования приведена на рис. 1.1.

2. Прогнозирование одномерных временных рядов

Временной ряд – это последовательность наблюдаемых данных через равные промежутки времени. Чтобы предсказать поведение временного ряда в будущем, необходимо построить модель, и если модель идентифицирована (т.е. подобрана определенного класса модель), то с ее помощью можно осуществить прогноз. Для построения модели необходимо произвести анализ временного ряда. Анализ временного ряда предполагает, что данные содержат несколько компонент: тренд, сезонную, циклическую и случайную ошибку. Большинство методов исследования временных рядов включает различные способы удаления случайной ошибки, чтобы выделить регулярную составляющую.

Тренд представляет собой общую систематическую линейную или нелинейную компоненту, которая плавно изменяется во времени и зависит от влияния долговременных факторов.

Сезонная составляющая – это периодически повторяющаяся компонента во времени, она служит источником краткосрочных колебаний временного ряда.

Циклическая компонента описывает длительные периоды подъемов и спадов. В экономических временных рядах это связано с тем, что экономическая активность не постоянна и имеет периоды подъемов и спадов.

Если временные ряды содержат значительную ошибку, то первым шагом выделения тренда является сглаживание.

Самый общий метод сглаживания – это метод скользящего среднего, в котором каждый член ряда заменяется средним из n соседних членов, где n – ширина уровня временного ряда (для квартальных данных $n = 4$, а для месячных уровней $n = 12$).

Метод среднего абсолютного прироста

Средний абсолютный прирост является обобщающим показателем скорости изменения ряда динамики во времени и дает возможность определить, насколько в среднем за единицу времени должен увеличиться (снизиться) уровень ряда в абсолютном выражении, чтобы от начального уровня за данное число периодов достигнуть конечного уровня. Для его определения используют формулу средней арифметической простой

$$\Delta Y_{mean} = \frac{\sum_{i=1}^{n-1} \Delta Y_i}{n-1} \text{ или } \Delta Y_{mean} = \frac{y_n - y_1}{n-1},$$

где $\Delta Y_i = Y_i - Y_{i-1};\ i = 2,3,..n;$ ΔY_{mean} – средний абсолютный прирост.

Прогноз по данному методу может быть осуществлен в том случае, если общая тенденция развития ряда динамики является линейной либо выполняется следующее условие: $S_e^2 \le \rho^2$, где S_e^2 – остаточная дисперсия, которая определяется по формуле

$S_e^2 = \frac{\sum_{i=2}^{n} (Y_i - \hat{Y}_i)^2}{n-2}$, тогда $\rho^2 = 0{,}5 \cdot \Delta Y_{mean}$; $\hat{Y}_i = Y_{i-1} + \Delta Y_{mean}$,

где $\hat{Y}_i$ – теоретическое значение ряда динамики.

Зная уровень ряда динамики Y_i, принятый за базу экстраполяции, можно записать экстраполяционную формулу для прогнозирования ряда динамики

$$\hat{Y}_{n+l} = Y_n + \Delta Y_{mean} \cdot l,$$

где l – период упреждения ($l = 1,2,\ldots$).

Метод среднего темпа роста

Дан ряд динамики $y_1, y_2, y_3, \ldots, y_n$, и пусть для этого ряда выполняются следующие условия $\tau_1 = \frac{y_2}{y_1} = \tau_2 = \frac{y_3}{y_2} = \tau_3 = \frac{y_4}{y_3} = \ldots, = \tau_{n-1} = \frac{y_n}{y_{n-1}}$, где τ – знаменатель геометрической прогрессии.

Тогда ряд динамики можно переписать следующим образом:

$y_1, y_1 \cdot \tau, y_1 \cdot \tau^2, \ldots y_1 \cdot \tau^{n-1}$.

Любой уровень ряда динамики можно выразить следующим образом:

$y_i = y_1 \cdot \tau^{i-1}$.

Из последнего уравнения выразим τ: $\tau = \sqrt[i-1]{\frac{y_i}{y_1}}$.

Так как темпы роста из года в год колеблятся, то для прогнозирования исчисляют его среднюю величину по формуле: $\bar{\tau} = \sqrt[n-1]{\frac{y_n}{y_1}}$.

Средний темп роста показывает, во сколько раз в среднем за единицу времени изменится уровень ряда динамики. Если в основу прогнозного расчета положить средний темп роста, то прогнозируемое значение ряда можно получить по формуле $\hat{y}_{n+1} = y_n \cdot \bar{\tau}^l$, где y_n – уровень ряда динамики

Метод линейного тренда (метод наименьших квадратов)

Закономерности развития явления во времени можно описать тем или иным видом функции $\hat{y} = f(x)$ и осуществить экстраполяции путем подстановки в функцию значения независимой переменной t. Экстраполяция дает возможность получить точечное значение прогноза, которое является средней оценкой для прогнозируемого интервала. Следует иметь в виду, что осуществлять прогноз можно только в том случае, если развитие явления достаточно хорошо описывается данным видом функции и условия,

определяющие тенденцию развития в прошлом, не претерпят существенных изменений в будущем.

Существуют несколько десятков математических функций, мы же рассмотрим только наиболее часто используемые для прогнозирования:

- линейная – $\hat{y} = a_0 + a_1 t$;
- параболическая – $\hat{y} = a_0 + a_1 t + a_2 t^2$;
- экспоненциальная – $\hat{y} = a_0 a_1^{\,t}$;
- полулогарифмическая – $\hat{y} = a_0 + a_1 \ln t$;
- гиперболическая – $\hat{y} = a_0 + \frac{a_1}{t}$;
- степенная – $\hat{y} = a_0 t^{a_1}$.

Полиномы невысоких степеней имеют конкретную интерпретацию рядов динамики. Например, для линейной и параболической функций коэффициенты модели можно интерпретировать следующим образом: a_0 – характеризует средние условия ряда динамики; a_1 – скорость роста ряда динамики; a_2 – ускорение ряда динамики.

Существует правило выбора степени полинома, основанное на определении величин конечных разностей уровней ряда динамики. Согласно этому правилу полином первой степени (линейная функция) применяется в том случае, если первые разности (абсолютные приросты) постоянны, полином второй степени – когда постоянны вторые разности и т.д.

Для выбора экспоненциальной функции необходима стабильность цепных темпов роста $Tp = y_i / y_{i-1} \approx const, i = 2,3,\ldots,n$.

Если возрастание одной величины приводит к убыванию другой, то используют гиперболическую или степенную функции, а если с возрастанием одной величины наблюдается замедленное возрастание другой, то используют полулогарифмическую функцию.

Модели сезонной декомпозиции

Идея сезонной декомпозиции состоит в том, что временной ряд представляют состоящим из четырех компонент:

- тренда (TR_t, где t обозначает момент времени);
- сезонной компоненты (SN_t);
- циклической компоненты (CL_t);
- случайной компоненты (IR_t).

Отличие между сезонной и циклической компонентой состоит в том, что циклическая компонента обычно имеет более длительный эффект, который к

тому же меняется от цикла к циклу, а сезонная имеет периодическую повторяемость через определенное время. В зависимости от взаимосвязи выделенных компонент различают модели аддитивные и мультипликативные.

Аддитивная модель: $y_t = TR_t + SN_t + CL_t + IR_t$.

Мультипликативная модель: $y_t = TR_t \cdot SN_t \cdot CL_t \cdot IR_t$, где y_t – значение временного ряда в момент времени t.

В каком случае проявляется аддитивная или мультипликативная сезонность? На графике это различие будет проявляться следующим образом: в аддитивном случае ряд будет иметь постоянные сезонные колебания, величина которых не зависит от общего уровня значений ряда; в мультипликативном случае величина сезонных колебаний будет меняться в зависимости от общего уровня значений ряда.

Так как циклическая компонента отличается от сезонной компоненты тем, что она обычно имеет большую временную протяженность и проявляется через неравные промежутки времени, то она также может быть либо аддитивной, либо мультипликативной как и в случае с сезонной компонентой.

Следовательно, аддитивная модель характеризуется тем, что сезонные и циклические колебания у нее остаются постоянными, а у мультипликативной модели сезонные и циклические колебания изменяются в зависимости от уровня ряда.

Прогнозные значения временного ряда определяются по формулам:

- аддитивная модель: $\hat{y}_t = TR_t + SN_t + CL_t$;

- мультипликативная модель: $\hat{y}_t = TR_t \cdot SN_t \cdot CL_t$.

Для практического изучения данного способа необходимо:

- сначала вычислить скользящие средние для временного ряда, при этом ширина сглаживания берется равной периоду сезонности. Если период сезонности – четное число, то необходимо полученные скользящие средние центрировать, чтобы отнести расчетные значения к определенной дате.

- после нахождения скользящих средних исключается сезонная изменчивость, в случае аддитивной модели ряд скользящих средних вычитается из наблюдаемого ряда или в случае для мультипликативной модели значения наблюдаемого ряда делятся на значения скользящих средних. Таким образом, получают ряд динамики с тренд-циклической компонентой (d_t).

- на последнем шаге выделяется случайная компонента путем вычитания из первоначального ряда тренд-циклической компоненты для аддитивной модели или делением этого ряда на тренд-циклическую компоненту для мультипликативной модели.

Адаптивные методы прогнозирования

Метод авторегрессии и скользящего среднего Бокса-Дженкинса

Важным классом вероятностных моделей являются линейные стационарные модели. Они основаны на предположении, что процесс остается в равновесии относительно постоянного среднего уровня. Но многие

временные ряды, описывающие экономические процессы, являются нестационарными и не колеблятся около фиксированного среднего.

Мы рассмотрим применение как стационарных моделей, так и не стационарных моделей, сводящихся к стационарным, взятием d -й разности наблюдаемого процесса или явления.

Рассмотрим некоторые понятия, которые будут использоваться в данном методе.

Оператор сдвига назад B, определяемый как $B \cdot y_t = y_{t-l}$, откуда по анологии получим $B^m y_t = y_{t-m}$, где m – сдвиг на m шагов.

Разностный оператор со сдвигом назад ∇y_t, который можно выразить, используя оператор сдвига назад B следующим образом

$\nabla y_t = y_t - y_{t-1} = (1-B) \cdot y_t$.

Модель авторегрессии (AR) – в этой модели текущее значение наблюдаемого ряда выражается как линейная совокупность предыдуших значений ряда и независимой от них случайной величиной e_t. В разные моменты времени t случайные величины e_t также независимы между собой и одинаково распределены, например, по нормальному закону.

Пусть имеем значение временного ряда в моменты времени t, $t-1$, $t-2$,… в виде y_t, y_{t-1}, y_{t-2}, … и пусть значения $\tilde{y}_t$, $\tilde{y}_{t-1}$, $\tilde{y}_{t-2}$,… – отклонения от среднего значения $\bar{y}_t$ данного временного ряда, т.е., например, $\tilde{y}_t = y_t - \bar{y}_t$, тогда процесс авторегрессии порядка p для стационарной модели можно записать следующим образом $\tilde{y}_t = \phi_1 \cdot \tilde{y}_{t-1} + \phi_2 \cdot \tilde{y}_{t-2} + \ldots + \phi_p \cdot \tilde{y}_{t-p} + e_t$, где $\phi_1, \phi_2, \ldots \phi_p$ – коэффициенты модели.

Определим оператор авторегрессии порядка p следующим образом

$\phi(B) = 1 - \phi_1 \cdot B - \phi_2 \cdot B^2 - \ldots - \phi_p \cdot B^p$,

тогда модель авторегрессии порядка p можно записать компактно $\phi(B) \cdot \tilde{y}_t = e_t$.

Для нестационарного ряда динамики авторегрессионная модель будет иметь вид $\phi(B) \cdot (1-B)^d \cdot y_t = e_t$.

Следовательно, нестационарный процесс может быть описан моделью авторегрессии, которая требует, чтобы d -я разность процесса была стационарной. Практически d равно 1 или 2.

Модель скользящего среднего (MA). Пусть значение $\tilde{y}_t$ линейно зависит от конечного числа q предыдущих значений et, такой процесс называют процессом скользящего среднего порядка q и для стационарных временных

рядов имеет вид $\tilde{y}_t = e_t - \theta_1 \cdot e_{t-1} - \theta_2 \cdot e_{t-2} - ... - \theta_q \cdot e_{t-q}$, где θ_1, θ_2 ,…, θ_q – коэффициенты модели.

Запишем оператор скользящего среднего порядка q следующим образом:

$\theta(B) = 1 - \theta_1 \cdot B - \theta_2 \cdot B^2 - ... - \theta_p \cdot B^q$.

В компактном виде модель такова: $\tilde{y}_t = \theta(B) \cdot e_t$.

Для нестационарного ряда динамики модель скользящего среднего будет иметь вид $(1-B)^d \cdot y_t = \theta(B) \cdot e_t$.

Смешанные модели авторегрессии и скользящего среднего (ARMA).

Для более гибкой подгонки моделей к наблюдаемым временным рядам возможно объединение моделей авторегрессии и скользящего среднего.

Стационарная модель: $\phi(B) \cdot \tilde{y}_t = \theta(B) \cdot e_t$.

Нестационарная модель: $\phi(B) \cdot (1-B)^d \cdot y_t = \theta(B) \cdot e_t$.

Представим данные модели в виде разностных уравнений, что позволит выразить текущее значение yt через их предшествующие значения и текущие и предшествующие значения e_t, e_{t-1}.

Стационарная модель:

$$(1 - \phi_1 \cdot B - \phi_2 \cdot B^2 - ... - \phi_p \cdot B^p) \cdot \tilde{y}_t = (1 - \theta_1 \cdot B - \theta_2 \cdot B^2 - ... - \theta_q \cdot B^q) \cdot e_t$$

$$\tilde{y}_t - = (1 - \theta_1 \cdot B - \theta_2 \cdot B^2 - ... - \theta_q \cdot B^q) \cdot e_t;$$

$$\tilde{y}_t - \phi_1 \cdot \tilde{y}_{t-1} - \phi_2 \cdot \tilde{y}_{t-2} - ... - \phi_p \cdot \tilde{y}_{t-p} = e_t - \theta_1 \cdot e_{t-1} - \theta_2 \cdot e_{t-2} - ... - \theta_q \cdot e_{t-q};$$

$$\tilde{y}_t = \phi_1 \cdot \tilde{y}_{t-1} + \phi_2 \cdot \tilde{y}_{t-2} + ... + \phi_p \cdot \tilde{y}_{t-p} + e_t - \theta_1 \cdot e_{t-1} - \theta_2 \cdot e_{t-2} - ... - \theta_q \cdot e_{t-q}.$$

Нестационарная модель:

$$(1 - \phi_1 \cdot B - \phi_2 \cdot B^2 - ... - \phi_p \cdot B^p) \cdot (1-B)^d y_t = (1 - \theta_1 \cdot B - \theta_2 \cdot B^2 - ... - \theta_q \cdot B^q) \cdot e_t$$

.

Введем обозначение $(1-B)^d y_t = w_t$, где w_t – стационарный временной ряд.

Пусть $p = 1, d = 2$ и $q = 2$, тогда нестационарная модель авторегрессии и скользящего среднего, обозначается ARMA (1, 2, 2), примет вид

$$(1 - \phi_1 \cdot B) \cdot (1 - B^2) \cdot y_t = (1 - \theta_1 \cdot B - \theta_2 \cdot B^2) \cdot e_t$$

$$(1 - B^2) \cdot y_t = (1 - 2B + B^2) \cdot y_t = y_t - 2y_{t-1} + y_{t-2}.$$

Подставим полученные значения вместо выражения $(1 - B^2) \cdot y_t$

$$(1 - \phi_1 \cdot B) \cdot (y_t - 2y_{t-1} + y_{t-2}) = (1 - \theta_1 \cdot B - \theta_2 \cdot B^2) \cdot e_t.$$

После преобразования получим

$$y_t = (2 + \phi_1) \cdot y_{t-1} - (1 + 2\phi_1) \cdot y_{t-2} + \phi_1 \cdot y_{t-3} + e_t - \theta_1 \cdot e_{t-1} - \theta_2 \cdot e_{t-2}.$$

В момент времени t значение $e_t = 0$.

Критерий стационарности – воспользуемся непараметрическим критерием, для проверки временного ряда на стационарность. Если в каком-либо ряду из n измерений, записанных в порядке их получения, за некоторым выбранным числом следует меньшее по величине число, то имеет место инверсия (I). Нулевая гипотеза состоит в том, что наблюдения y_t являются независимыми, т.е. тренд отсутствует. Используют двухсторонний критерий.

Идентификация параметров p и q.

Стационарные модели.

Для идентификации порядка модели используют графики, автокорреляционную функцию (АКФ) и частную автокорреляционную функцию (ЧАКФ). Большинство встречающихся на практике временных рядов можно с достаточной степенью точности аппроксимировать одной из пяти основных моделей, число параметров которых невелико – не больше двух:

- модель AR(1,0,0) – АКФ экспоненциально убывает, ЧАКФ имеет резко выделяющееся значение для первой задержки $k = 1$, а для остальных задержек выборочные ЧАКФ по абсолютной величине не превышают стандартную ошибку с выбранным уровнем значимости $\left|\hat{\phi}_{kk}\right| = t_\alpha / \sqrt{n}$;

- модель AR(2,0,0) – АКФ имеет форму синусоиды или экспоненциально убывает, ЧАКФ имеет резко выделяющееся значение для задержек $k = 1$ и 2, а для остальных задержек выборочные ЧАКФ по абсолютной величине не превышают стандартную ошибку с выбранным уровнем значимости;

- модель MA(0,0,1) – АКФ имеет резко выделяющееся значение при задержке $k = 1$, а для остальных задержек выборочные АКФ по абсолютной величине не превышают свою стандартную ошибку с выбранным уровнем значимости

$$\left|r_k\right| = t_\alpha \cdot \sqrt{\frac{1 + 2(r_1^2 + r_2^2 + ... + r_q^2)}{n}},$$

ЧАКФ экспоненциально убывает;

- модель MA(0,0,2) – АКФ имеет резко выделяющееся значение при задержках $k = 1$ и 2, а для остальных задержек выборочные АКФ по абсолютной величине не превышают свою стандартную ошибку с выбранным уровнем значимости. ЧАКФ имеет форму синусоиды или экспоненциально убывает;

- модель ARMA(1,0,1) – АКФ и ЧАКФ экспоненциально убывает при задержке $k = 1$.

Нестационарные модели.

Класс моделей, для которых d -я разность является стационарной (среднее постоянно, а выборочные дисперсия и автокорреляция не меняются во

времени) называется смешанным процессом авторегрессии и проинтегрированного скользящего среднего (ARIMA).

Берутся разности ряда столько раз, пока полученный ряд не станет стационарным. Количество разностей определяется параметром d . Для стабилизации дисперсии применяют логарифмическое преобразование.

После взятия первых разностей устраняют линейный тренд, а второе взятие разностей устраняет квадратичный тренд. На графике, в первом случае до взятия разностей, исходный ряд может иметь резкие колебания, которые имеют линейную тенденцию изменения без резких изменений наклона, если же в исходном ряде наблюдается резкое изменение наклона уровней ряда, то необходимо применить последовательно и взятие первых и вторых разностей. Если в исследуемом ряду присутствует сезонная компонента, то требуется взятие соответствующей сезонной разности.

Нестационарность процесса можно идентифицировать и с помощью коррелограммы.

Коррелограмма – это графическое представление выборочных коэффициентов автокорреляционной функции (АКФ) и их стандартных ошибок для последовательности задержек $k = 1,2,3,\ldots,n/4$. На графике приводится ширина интервала в размере двух стандартных ошибок при каждой задержке.

Нестационарность процесса можно идентифицировать по отсутствию быстрого спада выборочной АКФ и необязательно, чтобы АКФ были велики при малых задержках. Необходимый порядок разности считается достигнутым, если АКФ быстро затухает.

Приведем наиболее часто встречающие модели – ARIMA(0,1,1), ARIMA(1,1,0), ARIMA(1,1,1), ARIMA(0,2,2):

- модель ARIMA(1,d,0) – АКФ – экспоненциально затухает, ЧАКФ – только $\phi_{11} \neq 0$.

- модель ARIMA(0,d,1) – АКФ – только $r_1 \neq 0$, ЧАКФ – экспоненциально затухает.

- модель ARIMA(2,d,0) – АКФ – имеет форму затухающей синусоиды или экспоненциально убывает. ЧАКФ – только $\phi_{11} \neq 0$ и $\phi_{22} \neq 0$.

- модель ARIMA(0,d,2) – АКФ – только $r_1 \neq 0$ и $r_2 \neq 0$, ЧАКФ – имеет форму затухающей синусоиды или экспоненциально убывает.

- модель ARIMA(1,d,1) – АКФ и ЧАКФ – экспоненциально затухают при $k = 1$.

- модель ARIMA(p,d,q)(ps,ds,qs) – мультипликативная сезонная модель авторегрессии и проинтегрированного скользящего среднего.

3. Эконометрические модели прогнозирования

Возможны другие модели прогноза – причинно-следственные (каузальные), когда модель можно представить, связывая одну или несколько переменных, которые являются причиной изменения другой (зависимой) переменной. Поэтому выделяют однофакторные и многофакторные модели прогнозирования. Прогноз получают подстановкой независимой или независимых переменных и в данном случае прогноз может быть содержательно интерпретирован.

Оценивание параметров модели $y_i = a_0 + a_1 x_i + e_i \quad (i = 1,\ldots,n)$ с помощью МНК $SSE = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 \to \min$ требует выполнения определенных условий:

- $E(e_i)$ – математическое ожидание случайных ошибок (e_i) имеет среднюю равную нулю;

- $E(e_i^2) = \sigma^2$ – для любого заданного x_i распределение y_{ij} имеет постоянную дисперсию (это условие называется гомоскедастичностью, если это условие не выполняется – называется гетероскедастичностью и тогда необходимо применять взвешенный МНК), но для экономических данных y_{ij}, $j = 1$, и мы предполагаем, что значение y_i извлекается из генеральной совокупности y_{ij};

- $E(e_i e_j)$ – математическое ожидание ковариации ошибок (автокорреляций) $(e_i e_j)$ равно нулю при $i \neq j$.

Если перечисленные условия выполняются, то оценки параметров модели обладают следующими свойствами:

- несмешанностью – т.е. математическое ожидание оценки параметров модели равны их истинным значениям $E(a_0) = \alpha_0$ и $E(a_1) = \alpha_1$;

- сосоятельностью – дисперсия оценок параметров при $n \to \infty$ равна нулю;

- эффективностью – параметры модели имеют наименьшую дисперсию по сравнению с другими оценками.

Многофакторные модели

Оценка параметров модели при $m > 1$ (m – количество независимых переменных) остается такой же, как и для однофакторных моделей $\mathbf{a} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$, но определения $\mathbf{a}$ и $\mathbf{X}$ принимают следующий вид

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ 1 & x_{31} & x_{32} & \cdots & x_{3m} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}; \quad \mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix},$$

где $\mathbf{X}$ – матрица наблюдаемых значений независимых переменных; $\mathbf{a}$ – вектор оценок параметров модели; n – количество наблюдаемых значений независимых переменных.

Проверка параметров модели на значимость, определение стандартных ошибок параметров модели и самого уравнения, а также стандартная оценка ошибки прогноза выполняются по тем же формулам, что и для однофакторной модели.

Линейную многофакторную модель можно представить в следующем виде

$$\hat{y} = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_m x_m .$$

На данную модель накладываются те же ограничения, что и для модели с одной независимой переменной, но с добавлением четвертого условия – ранг матрицы $\mathbf{X}$ равен m и $m < n$. Это условие означает линейную независимость векторов матрицы независимых переменных, и эквивалентно тому, что матрица (XTX)-1 должна иметь полный ранг. При нарушении этого условия, т.е., когда один из столбцов матрицы $\mathbf{X}$ является линейной комбинацией остальных столбцов, такая матрица называется вырожденной и ее определитель равен нулю.

На практике чаще бывает такая ситуация, когда матрица независимых переменных $\mathbf{X}$ имеет полный ранг, но между независимыми переменными существует высокая корреляция, то есть матрица XTX близка к вырожденной и ее определитель мал. Определенные по такой обратной матрице стандартные ошибки коэффициентов регрессии велики, и оценка их значимости по t - критерию Стьюдента может привести к неверным выводам о незначимости включения той или иной независимой переменной в модель, что приводит к ошибке спецификации. Небольшие изменения в исходных данных приводят к существенному изменению оценок коэффициентов модели. Эти коэффициенты могут иметь другую направленность воздействия на зависимую переменную по сравнению с теоретической. Такую проблему называют мультиколлинеарностью.

Феррaром и Глобером предложен критерий оценки мультиколлинеарности, суть которого в следующем. Если независимые переменные x_1, x_2,…, x_m попарно линейно независимы, то корреляционная матрица является единичной, следовательно, матрица, обратная к корреляционной матрице, также является единичной. Выборочная

корреляционная матрица и ее обратная отличны от единичной. На этом свойстве основан предложенный ими критерий, который подчиняется распределению Фишера с $v_1 = n - m$; $v_2 = m - 1$ степенями свободы

$$F = \left|c_{ii} - 1\right| \cdot \frac{n-m}{m-1},$$

где c_{ii} – диагональные элементы обратной корреляционной матрицы независимых переменных; m – количество независимых переменных; n – количество наблюдений.

Если $F < F_{кр}$; $\alpha = 0{,}05$; v_1; v_2 то принимается гипотеза отсутствия мультиколлинеарности j независимой переменной с остальными независимыми переменными при уровне значимости $\alpha = 0{,}05$.

Используется также критерий обнаружения мультиколлинеарности, основанный на множественном коэффициенте детерминации (R^2) между независимой переменной и остальными независимыми переменными $(x_j$ и x_1, $x_{j-1}, \ldots, x_m)$, определенный как VIF (Variance Inflation Factor): $VIF_j = \frac{1}{1 - R_j^2}$.

Если $R_j^2 = 0$, то $VIF_j = 1$ и x_j не зависит от остальных независимых переменных, если $R_j^2 > 0$, то x_j зависит от остальных независимых переменных и $VIF_j > 1$. Определив VIF для каждой независимой переменной, получим, $VIF \geq 5$ – это свидетельствует о наличии мультиколлинеарности между x_j и остальными независимыми переменными для малых выборок, а если $VIF \geq 10$ – то для больших выборок.

Наличие мультиколлинеарности контролируется также величиной обусловленности (cond) матрицы парных коэффициентов корреляции независимых переменных, определяемой как отношение максимального собственного значения данной матрицы к минимальному собственному значению. Если $cond < 100$, то это свидетельствует об отсутствии проблемы мультиколлинеарности.

Модель, содержащая несколько уравнений

При статистическом моделировании сложных экономических процессов одна и та же переменная может в одном уравнении быть независимой переменной, т.е. являться причиной изменения результирующего показателя, а в другом сама выступает в роли зависимой переменной*. Рассмотрим простой пример, содержащий два уравнения, записанных в векторной форме

y1=α1 y2+α2 X1+ε1

y2=β1 y1+β2 X1+ε2

Если рассматривать каждое уравнение в отдельности, то они являются обычными регрессионными уравнениями и параметры модели α_1, α_2 и β_1, β_2

определяются по МНК. Но y1 и y2 входят как в правые так и в левые части уравнений и их уже нельзя считать независимыми, так как они определяются в один и тот же момент времени t.

Рассмотренный тип модели не позволяет делить переменные, входящие в модель, на зависимые и независимые, как это принято в регрессионных моделях с одним уравнением.

4. Динамические модели прогнозирования

В динамических задачах модели отражают зависимость входящих в них переменных не только от фактора времени, но и их временную взаимосвязь.

Принципы построения непрерывных динамических моделей

Исследование экономических процессов приводит к построению математических моделей, основой которых являются дифференциальные уравнения. Уравнение, связывающее независимую переменную, ее функции и производные этой функции называют дифференциальным уравнением. Дифференциальные уравнения делят на две группы. К первой группе относятся уравнения, в которые входят производные только по одной переменной. Такие уравнения называются обыкновенными дифференциальными уравнениями. Если в уравнение входят производные по нескольким переменным, то такие уравнения называются дифференциальными уравнениями в частных производных.

Принципы построения конечно-разностных динамических моделей (конкретный пример)

Для нефтяных месторождений нет общего рецепта выбора системы разработки, так как эта система зависит от многих факторов: размера залежи, мощности пласта, глубины залегания, характеристик нефти, содержание в ней газа и воды и др. Целью регулирования процесса разработки нефтяной залежи является равномерное перемещение водонефтяного контакта. При неравномерном продвижении воды в отдельных направлениях скважины могут быстро обводняться, поэтому принимаются меры по снижению обводненности скважин, например, такие как: ограничение отбора жидкости, проведение изоляционных работ, ограничение закачки воды в близкорасположенные нагнетательные скважины и др. Прогнозирование состояния обводненности скважин является важной задачей в процессе разработки нетяной залежи. В процессе разработки нефтяной залежи строят карты отборов нефти и карты обводненности по скважинам. Сравнение таких карт, построенных на разные даты времени, позволяют установить изменения указанных показателей по скважинам. Для оперативного прогноза можно построить модель обводненности скважины по статистическим данным и использовать ее для прогноза и регулирования обводненности.

Пусть в момент времени t доля воды по скважине составляет D_{B_t}, а доля нефти в этот же момент времени составит ($1-D_{B_t}$). За период времени Δt доля нефти уменьшится на величину коэффициента падения добычи нефти $\alpha = const (0<\alpha<1)$ и составит $\alpha(1-D_{B_t})$. Если в это же время проведено мероприятие по снижению обводненности скважины, то обводненность скважины можно определить так $\beta \cdot D_{B_t}$, где β коэффициент, характеризующий эффективность мероприятий, направленных на снижение обводненности скважин $(0<\beta<1$ и $\beta<\alpha)$.

Таким образом, изменение уровня обводненности скважины можно записать

$$\Delta D_{B_t} = D_{B_{t+1}} - D_{B_t} = \alpha \cdot (1 - D_{B_t}) - \beta \cdot D_{B_t}.$$

Выразим обводненность скважины в момент времени $D_{B_{t+1}}$

$$D_{B_{t+1}} = \alpha \cdot (1 - D_{B_t}) - \beta \cdot D + D_{B_t} = \alpha + (1-\alpha-\beta) \cdot D_{B_t} = \alpha + \gamma \cdot D_{B_t}$$

Пусть в момент времени $t=0$ известна доля обводненности D_{B_0}, тогда уравнение можно переписать следующим образом:

$$D_{B_1} = \alpha + \gamma D_{B_0}$$

$$D_{B_2} = \alpha + \gamma D_{B_1} = \alpha + \gamma(\alpha + \gamma D_{B_0}) = \alpha + \alpha\gamma + \gamma^2 D_{B_0}$$

– –

$$D_{B_t} = \alpha + \gamma D_{B_{t-1}} = \alpha + \alpha\gamma + \alpha\gamma^2 + \ldots + \alpha\gamma^{t-1} + \gamma^t D_{B_0} =$$

$$= \alpha(1 + \gamma + \gamma^2 + \ldots + \gamma^{t-1}) + \gamma^t D_{B_0}.$$

Полученное выражение $1+\gamma+\gamma^2+\ldots+\gamma^{t-1}$ представляет собой сумму геометрической прогрессии – $S_n = \frac{1(1-\gamma^t)}{1-\gamma}$, следовательно, обводненность скважины можно представить так:

$$D_{B_t} = \frac{\alpha(1-\gamma^t)}{1-\gamma} + \gamma^t D_{B_0} = \frac{\alpha}{\alpha+\beta} + (D_{B_0} - \frac{\alpha}{\alpha+\beta})(1-\alpha-\beta)^t .$$

Если не проводить регулирование процесса эксплуатации отдельных скважин, то скважины обводняются более высокими темпами и этот процесс можно выразить следующей формулой, учитывая, что $\beta = 0$

$$D_{B_t} = 1 + (D_{B_0} - 1)(1-\alpha)^t .$$

Рассмотренные модели являются практическим приложением математических методов исследования явлений и процессов.

Вопросы по лекции

1. На основе каких признаков можно классифицировать методы прогнозирования?
2. Назовите виды экстраполяции. В чем разница между экстраполяцией и интерполяцией?
3. Опишите основные компоненты временного ряда.
4. Опишите сущность метода наименьших квадратов.
5. Правила выбора параметров сглаживания.
6. Отличие методов экспоненциального сглаживания и вероятностного моделирования.
7. Показатели статистической обработки экспертных мнений.
8. Описание последовательности процесса идентификации моделей прогнозирования, содержащих сезонную компоненту?
9. Процедура построения прогнозов на базе модели ARIMA.
10. Описание модели Хольта
11. Описание модели Тейлора-Вейджа.
12. Описание метода МГУА

Лекция № 6 Нейросетевое прогнозирование

План лекции

1. Введение в теорию нейронных сетей
2. Основные модели нейронных сетей
3.Общий подход к прогнозированию с помощью нейронных сетей

1. Введение в теорию нейронных сетей

Традиционно используемым для описания нейронных сетей математическим языком является аппарат векторной и матричной алгебры. Для максимального упрощения изложения, ограничивая набор общематематических сведений только этим аппаратом, хотелось бы подчеркнуть, что в современной нейронауке широко используются и другие разделы математики. Среди них - дифференциальные уравнения, применяемые для анализа нейронных сетей в непрерывном времени, а также для построения детальных моделей нейрона; Фурье-анализ для описания поведения системы при кодировании в частотной области; теория оптимизации как основа для разработки алгоритмов обучения; математическая логика и булева алгебра - для описания двоичных сетей, и другие.

Элементом клеточной структуры мозга является нервная клетка - нейрон. Нейрон в своем строении имеет много общих черт с другими клетками биоткани: тело нейрона окружено плазматической мембраной, внутри которой находится цитоплазма, ядро и другие составляющие клетки. Однако нервная клетка существенно отличается от иных по своему функциональному назначению. Нейрон выполняет прием, элементарное преобразование и дальнейшую передачу информации другим нейронам. Информация переносится в виде импульсов нервной активности, имеющих электрохимическую природу.

Нейроны крайне разнообразны по форме, которая зависит от их местонахождения в нервной системе и особенностей функционирования. На Рис. 3.1. приведена схема строения "типичного" нейрона. Тело клетки содержит множество ветвящихся отростков двух типов. Отростки первого типа, называемые дендритами за их сходство с кроной раскидистого дерева, служат в качестве входных каналов для нервных импульсов от других нейронов. Эти импульсы поступают в сому или тело клетки размером от 3 до 100 микрон, вызывая ее специфическое возбуждение, которое затем распространяется по выводному отростку второго типа - аксону. Длина аксонов обычно заметно превосходит размеры дентритов, в отдельных случаях достигая десятков сантиметров и даже метров.

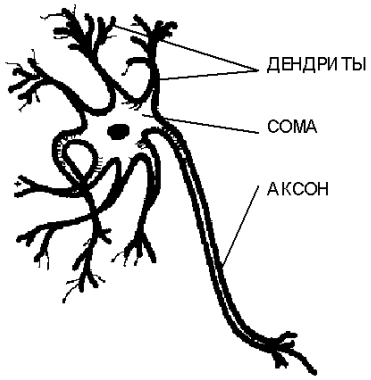


Рис. 6.1 Общая схема строения биологического нейрона.

Взаимодействующие между собой посредством передачи через отростки возбуждений нейроны формируют нейронные сети. Переход от рассмотрения отдельного нейрона к изучению нейронных сетей является естественным шагом в нейробиологической иерархии.

Общее число нейронов в центральной нервной системе человека достигает 1010 - 1011, при этом каждая нервная клетка связана в среднем с 103 - 104 других нейронов. Установлено, что в головном мозге совокупность нейронов в объеме масштаба 1 мм3формирует относительно независимую локальную сеть, несущую определенную функциональную нагрузку.

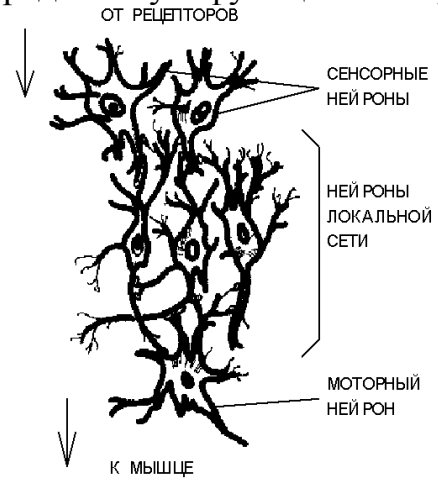


Рис. 6.2.Структура простой рефлекторной нейронной сети.

Исторически первой работой, заложившей теоретический фундамент для создания искусственных моделей нейронов и нейронных сетей, принято считать опубликованную в 1943 г. статью Уоррена С.МакКаллока и Вальтера Питтса "Логическое исчислени идей, относящихся к нервной активности". Главный принцип теории Маккалока и Питтса заключается в том, что произвольные явления, относящиеся к высшей нервной деятельности, могут быть проанализированы и поняты, как некоторая активность в сети, состоящей из логических элементов, принимающих только два состояния ("все или ничего"). При этом для всякого логического выражения, удовлетворяющего указанным авторами условиям, может быть найдена сеть логических элементов, имеющая описываемое этим выражением поведение.

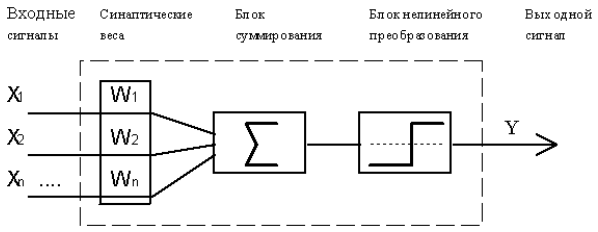


Рис. 6.3 Функциональная схема формального нейрона Маккалока и Пиитса.

В блоке суммирования происходит накопление общего входного сигнала (обычно обозначаемого символом net), равного взвешенной сумме входов:

$$net = \sum_{i=1}^{n} W_i x_i$$

Опишем наиболее широко используемые типы переходных функций Y=f(net).

Пороговая функция (рассмотренная Маккалоком и Питтсом):

$$Y = f(net) = \begin{cases} 1, & net > \Theta \\ 0, & net \le \Theta \end{cases}$$

Линейная функция, а также ее вариант - линейная функция с погашением отрицательных сигналов:

$$Y = f(net) = \begin{cases} net, & net > \Theta \\ 0, & net \le \Theta \end{cases}$$

Сигмоидальная функция:

$$Y = f(net) = \frac{1}{1 + \exp(-(net - \Theta))}$$

Как указывалось еще С.Гроссбергом, сигмоидальная функция обладает избирательной чувствительностью к сигналам разной интенсивности, что соответствует биологическим данным. Наибольшая чувствительность наблюдается вблизи порога, где малые изменения сигнала net приводят к ощутимым изменениям выхода. Напротив, к вариациям сигнала в областях значительно выше или ниже порогового уровня сигмоидальная функция не чувствительна, так как ее производная при больших и малых аргументах стремится к нулю.

2. Основные модели нейронных сетей

Персептрон Розенблатта.

Одной из первых искусственных сетей, способных к перцепции (восприятию) и формированию реакции на воспринятый стимул, явился PERCEPTRON Розенблатта (F.Rosenblatt, 1957). Персептрон рассматривался его автором не как конкретное техническое вычислительное устройство, а как модель работы мозга. Нужно заметить, что после нескольких десятилетий исследований современные работы по искусственным нейронным сетям редко преследуют такую цель.

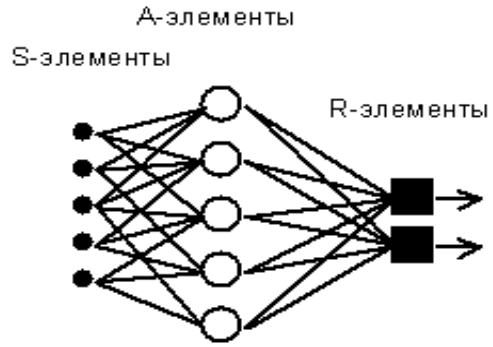


Рис. 6.4 Элементарный персептрон Розенблатта.

Простейший классический персептрон содержит нейроподобные элементы трех типов, назначение которых в целом соответствует нейронам рефлекторной нейронной сети, рассмотренной в предыдущей лекции. S-элементы формируют сетчатку сенсорных клеток, принимающих двоичные сигналы от внешнего мира. Далее сигналы поступают в слой ассоциативных или А-элементов (для упрощения изображения часть связей от входных S-клеток к А-клеткам не показана). Только ассоциативные элементы, представляющие собой формальные нейроны, выполняют нелинейную обработку информации и имеют изменяемые веса связей. R-элементы с фиксированными весами формируют сигнал реакции персептрона на входной стимул.

По своей организации и функциональному назначению искусственная нейронная сеть с несколькими входами и выходами выполняет некоторое преобразование входных стимулов - сенсорной информации о внешнем мире - в выходные управляющие сигналы. Число преобразуемых стимулов равно n - числу входов сети, а число выходных сигналов соответствует числу выходов m. Совокупность всевозможных входных векторов размерности n образует векторное пространство X, которое мы будем называть признаковым пространством Аналогично, выходные вектора также формируют признаковое пространство, которое будет обозначаться Y. Теперь нейронную сеть можно мыслить, как некоторую многомерную функцию $F: X \rightarrow Y$, аргумент которой принадлежит признаковому пространству входов, а значение - выходному признаковому пространству.

При произвольном значении синаптических весовых коэффициентов нейронов сети функция, реализуемая сетью также произвольна. Для получения требуемой функции необходим специфический выбор весов. Упорядоченная совокупность всех весовых коэффициентов всех нейронов может быть представлена, как вектор W. Множество всех таких векторов также формирует векторное пространство, называемое пространством состояний или конфигурационным (фазовым) пространством W. Термин "фазовое пространство" пришел из статистической физики систем многих частиц, где под ним понимается совокупность координат и импульсов всех частиц, составляющих систему.

Постановка задачи оптимизации при обучении нейронной сети

Пусть имеется нейронная сеть, выполняющая преобразование $F:X \rightarrow Y$ векторов X из признакового пространства входов X в вектора Y выходного

пространства Y. Сеть находится в состоянии W из пространства состояний W. Пусть далее имеется обучающая выборка (Xi,Yi), α = 1..p. Рассмотрим полную ошибку E, делаемую сетью в состоянии W.

$$E = E(W) = \sum_{\alpha} \left\| F(X^{\alpha};W) - Y^{\alpha} \right\| = \sum_{\alpha} \sum_{i} [F_i(X^{\alpha};W) - Y_i^{\alpha}]^2$$

Отметим два свойства полной ошибки. Во-первых, ошибка E=E(W) является функцией состояния W, определенной на пространстве состояний. По определению, она принимает неотрицательные значения. Во-вторых, в некотором обученном состоянии W*, в котором сеть не делает ошибок на обучающей выборке, данная функция принимает нулевое значение. Следовательно, обученные состояния являются точками минимума введенной функции E(W).

Таким образом, задача обучения нейронной сети является задачей поиска минимума функции ошибки в пространстве состояний, и, следовательно, для ее решения могут применяться стандартные методы теории оптимизации. Эта задача относится к классу многофакторных задач, так, например, для однослойного персептрона с N входами и M выходами речь идет о поиске минимума в NxM-мерном пространстве.

Многослойный ПЕРСЕПТРОН.

Рассмотрим иерархическую сетевую структуру, в которой связанные между собой нейроны (узлы сети) объединены в несколько слоев (Рис. 6.1). На возможность построения таких архитектур указал еще Ф.Розенблатт, однако им не была решена проблема обучения. Межнейронные синоптические связи сети устроены таким образом, что каждый нейрон на данном уровне иерархии принимает и обрабатывает сигналы от каждого нейрона более низкого уровня. Таким образом, в данной сети имеется выделенное направление распространения нейроимпульсов - от входного слоя через один (или несколько) скрытых слоев к выходному слою нейронов. Нейросеть такой топологии мы будем называть обобщенным многослойным персептроном или, если это не будет вызывать недоразумений, просто персептроном.

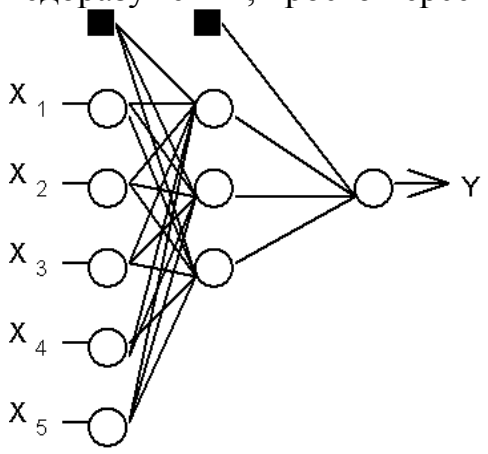


Рис.6.5 Структура многослойного персептрона с пятью входами, тремя нейронами в скрытом слое, и одним нейроном выходного слоя.

Персептрон представляет собой сеть, состоящую из нескольких последовательно соединенных слоев формальных нейронов МакКаллока и Питтса. На низшем уровне иерархии находится входной слой, состоящий из

сенсорных элементов, задачей которого является только прием и распространение по сети входной информации. Далее имеются один или, реже, несколько скрытых слоев. Каждый нейрон на скрытом слое имеет несколько входов, соединенных с выходами нейронов предыдущего слоя или непосредственно со входными сенсорами X1..Xn, и один выход. Нейрон характеризуется уникальным вектором весовых коэффициентов w. Веса всех нейронов слоя формируют матрицу, которую мы будем обозначать V или W. Функция нейрона состоит в вычислении взвешенной суммы его входов с дальнейшим нелинейным преобразованием ее в выходной сигнал:

$$y = 1 \Big/ \left(1 + \exp(-[\sum_i W_i x_i - \Theta]) \right)$$

Выходы нейронов последнего, выходного, слоя описывают результат классификации Y=Y(X). Особенности работы персептрона состоят в следующем. Каждый нейрон суммирует поступающие к нему сигналы от нейронов предыдущего уровня иерархии с весами, определяемыми состояниями синапсов, и формирует ответный сигнал (переходит в возбужденное состояние), если полученная сумма выше порогового значения. Персептрон переводит входной образ, определяющий степени возбуждения нейронов самого нижнего уровня иерархии, в выходной образ, определяемый нейронами самого верхнего уровня. Число последних, обычно, сравнительно невелико. Состояние возбуждения нейрона на верхнем уровне говорит о принадлежности входного образа к той или иной категории.

Традиционно рассматривается аналоговая логика, при которой допустимые состояния синаптических связей определяются произвольными действительными числами, а степени активности нейронов - действительными числами между 0 и 1. Иногда исследуются также модели с дискретной арифметикой, в которой синапс характеризуется двумя булевыми переменными: активностью (0 или 1) и полярностью (-1 или +1), что соответствует трехзначной логике. Состояния нейронов могут при этом описываться одной булевой переменной. Данный дискретный подход делает конфигурационное пространство состояний нейронной сети конечным (не говоря уже о преимуществах при аппаратной реализации).

Обучение методом обратного распространения ошибок

Для упрощения обозначений ограничимся ситуацией, когда сеть имеет только один скрытый слой. Матрицу весовых коэффициентов от входов к скрытому слою обозначим W, а матрицу весов, соединяющих скрытый и выходной слой - как V. Для индексов примем следующие обозначения: входы будем нумеровать только индексом i, элементы скрытого слоя - индексом j, а выходы, соответственно, индексом k.

Пусть сеть обучается на выборке (X_i,Y_i), i =1..p. Активности нейронов будем обозначать малыми буквами y с соответствующим индексом, а суммарные взвешенные входы нейронов - малыми буквами x.

Таблица 6.1. Алгоритм обратного распространения ошибки.

Шаг 0.	Начальные значения весов всех нейронов всех слоев V(t=0) и W(t=0) полагаются случайными числами.

Шаг 1.	Сети предъявляется входной образ X□, в результате формируется выходной образ y□ Y□. При этом нейроны последовательно от слоя к слою функционируют по следующим формулам: скрытый слой $$x_j = \sum_i W_{ij} X_i^{\alpha}; y_j = f(x_j)$$ выходной слой $$x_k = \sum_j V_{jk} y_j; y_k = f(x_k)$$ Здесь f(x) - сигмоидальная функция, определяемая по формуле (6.1)
Шаг 2.	Функционал квадратичной ошибки сети для данного входного образа имеет вид: $$E = 1 / 2 \sum_k \left(y_k - Y_k^{\alpha}\right)^2$$ Данный функционал подлежит минимизации. Классический градиентный метод оптимизации состоит в итерационном уточнении аргумента согласно формуле: $$V_{jk}(t+1) = V_{jk}(t) - h \cdot \frac{\partial E}{\partial V_{jk}}$$ Функция ошибки в явном виде не содержит зависимости от веса Vjk, поэтому воспользуемся формулами неявного дифференцирования сложной функции: $$\frac{\partial E}{\partial y_k} = \delta_k = \left(y_k - Y_k^{\alpha}\right);$$ $$\frac{\partial E}{\partial x_k} = \frac{\partial E}{\partial y_k} \cdot \frac{\partial y_k}{\partial x_k} = \delta_k \cdot y_k(1 - y_k)$$ $$\frac{\partial E}{\partial V_{jk}} = \frac{\partial E}{\partial y_k} \cdot \frac{\partial y_k}{\partial x_k} \cdot \frac{\partial x_k}{\partial V_{jk}} =$$ $$= \delta_k \cdot y_k(1 - y_k) \cdot y_j$$ Здесь учтено полезное свойство сигмоидальной функции f(x): ее производная выражается только через само значение функции, f'(x)=f(1-f). Таким образом, все необходимые величины для подстройки весов выходного слоя V получены.
Шаг 3.	На этом шаге выполняется подстройка весов скрытого слоя. Градиентный метод по-прежнему дает: $$W_{ij}(t+1) = W_{ij}(t) - h \cdot \frac{\partial E}{\partial W_{ij}}$$ Вычисления производных выполняются по тем же формулам, за исключением некоторого усложнения формулы для ошибки □ j. $$\frac{\partial E}{\partial x_k} = \frac{\partial E}{\partial y_k} \cdot \frac{\partial y_k}{\partial x_k} = \delta_k \cdot y_k(1 - y_k)$$ $$\frac{\partial E}{\partial y_j} = \delta_j = \sum_k \frac{\partial E}{\partial x_k} \cdot \frac{\partial x_k}{\partial y_j} =$$ $$= \sum_k \delta_k \cdot y_k(1 - y_k) \cdot V_{jk};$$ $$\frac{\partial E}{\partial W_{ij}} = \frac{\partial E}{\partial y_j} \cdot \frac{\partial y_j}{\partial x_j} \cdot \frac{\partial x_j}{\partial W_{ij}} =$$ $$= \delta_j \cdot y_j(1 - y_j) \cdot X_i^{\alpha} =$$ $$= \left[\sum_k \delta_k \cdot y_k(1 - y_k) \cdot V_{jk}\right] \cdot \left[y_j(1 - y_j) \cdot X_i^{\alpha}\right]$$ При вычислении X_j здесь и был применен принцип обратного распространения ошибки: частные производные берутся только по переменным последующего слоя. По полученным формулам модифицируются веса нейронов скрытого слоя. Если в нейронной сети имеется несколько скрытых слоев, процедура

	обратного распространения применяется последовательно для каждого из них, начиная со слоя, предшествующего выходному, и далее до слоя, следующего за входным. При этом формулы сохраняют свой вид с заменой элементов выходного слоя на элементы соответствующего скрытого слоя.
Шаг 4.	Шаги 1-3 повторяются для всех обучающих векторов. Обучение завершается по достижении малой полной ошибки или максимально допустимого числа итераций, как и в методе обучения Розенблатта.

Звезды Гроссберга

Нейрон в форме входной звезды имеет N входов X1..XN, которым соответствуют веса W1..XN, и один выход Y, являющийся взвешенной суммой входов. Входная звезда обучается выдавать сигнал на выходе всякий раз, когда на входы поступает определенный вектор. Таким образом, входная звезда является детектором совокупного состояния своих входов. Процесс обучения представляется в следующей итерационной форме:

$$W_i(t+1) = W_i(t) + \alpha\left(X_i - W_i(t)\right)$$

Темп обучения имеет начальное значение масштаба 0.1 и постепенно уменьшается в процессе обучения. В процессе настройки нейрон учится усредненным обучающим векторам.

Выходная звезда Гроссберга выполняет противоположную функцию - функцию командного нейрона, выдавая на выходах определенный вектор при поступлении сигнала на вход. Нейрон этого типа имеет один вход и M выходов с весами W1..M, которые обучаются по формуле:

$$W_i(t+1) = W_i(t) + \beta\left(Y_i - W_i(t)\right)$$

Рекомендуется начать с 0 порядка единицы и постепенно уменьшать до нуля в процессе обучения. Итерационный процесс будет сходиться к собирательному образу, полученному из совокупности обучающих векторов.

Особенностью нейронов в форме звезд Гроссберга является локальность памяти. Каждый нейрон в форме входной звезды помнит "свой" относящийся к нему образ и игнорирует остальные. Каждой выходной звезде присуща также конкретная командная функция. Образ памяти связывается с определенным нейроном, а не возникает вследствие взаимодействия множества нейронов в сети.

Карта самоорганизации Кохонена.

В противоположность хеммингговой сети модель Кохонена (T.Kohonen, 1982) выполняет обобщение предъявляемой информации. В результате работы НС Кохонена получается образ, представляющий собой карту распределения векторов из обучающей выборки. Таким образов, в модели Кохонена выполняется решение задачи нахождения кластеров в пространстве входных образов.

Данная сеть обучается без учителя на основе самоорганизации. По мере обучении вектора весов нейронов стремятся к центрам кластеров - групп векторов обучающей выборки. На этапе решения информационных задач сеть относит новый предъявленный образ к одному из сформированных кластеров, указывая тем самым категорию, к которой он принадлежит.

Рассмотрим архитектуру НС Кохонена и правила обучения подробнее. Сеть Кохонена, также как и сеть Липпмана-Хемминга, состоит из одного слоя нейронов. Число входов каждого нейрона равно размерности входного образа. Количество же нейронов определяется той степенью подробности с которой требуется выполнить кластеризацию набора библиотечных образов. При достаточном количестве нейронов и удачных параметрах обучения НС Кохонена может не только выделить основные группы образов, но и установить "тонкую структуру" полученных кластеров. При этом близким входным образам будет соответствовать близкие карты нейронной активности.

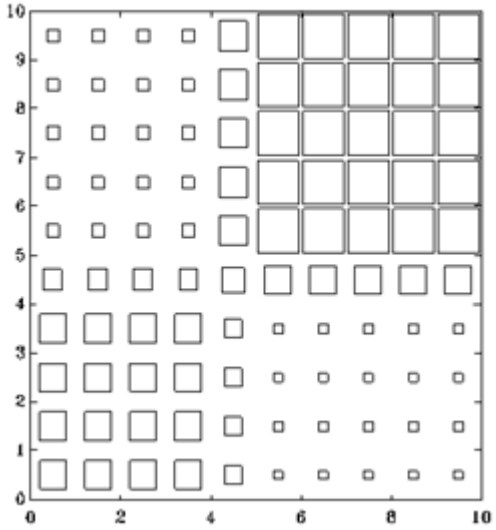


Рис. 6.6 Пример карты Кохонена. Размер каждого квадратика соответствует степени возбуждения соответствующего нейрона.

Обучение начинается с задания случайных значений матрице связей W_n^m. В дальнейшем происходит процесс самоорганизации, состоящий в модификации весов при предъявлении на вход векторов обучающей выборки. Для каждого нейрона можно определить его расстояние до вектора входа:

$$d_m = \sum_{i=1}^{N} \left(x_i(t) - W_i^m(t)\right)^2$$

Далее выбирается нейрон m=m*, для которого это расстояние минимально. На текущем шаге обучения t будут модифицироваться только веса нейронов из окрестности нейрона m*:

$$W_n^m(t+1) = W_n^m(t) + \eta\left(x_n(t) - W_n^m\right)$$

Первоначально в окрестности любого из нейронов находятся все нейроны сети, в последствии эта окрестность сужается. В конце этапа обучения подстраиваются только веса самого ближайшего нейрона. Темп обучения $\eta(t)<1$ с течением времени также уменьшается. Образы обучающей выборки предъявляются последовательно, и каждый раз происходит подстройка весов. Нейронная сеть Кохонена может обучаться и на искаженных версиях входных векторов, в процессе обучения искажения, если они не носят систематический характер, сглаживаются.

3.Общий подход к прогнозированию с помощью нейронных сетей

На НС задача прогнозирования формализуется через задачу распознавания образов. Данных о прогнозируемой переменной за некоторый промежуток времени образуют образ, класс которого определяется значением прогнозируемой переменной в некоторый момент времени за пределами данного промежутка т.е. значением переменной через интервал прогнозирования. Метод окон предполагает использование двух окон Wi и Wo с фиксированными размерами n и m соответственно. Эти окна, способны перемещаться с некоторым шагом по временной последовательности исторических данных, начиная с первого элемента, и предназначены для доступа к данным временного ряда, причем первое окно Wi, получив такие данные, передает их на вход нейронной сети, а второе - Wo - на выход. Получающаяся на каждом шаге пара Wi -> Wo (3.1) используется как элемент обучающей выборки (распознаваемый образ, или наблюдение) .

Например, пусть есть некоторые данные по состоянию объекта (k = 16) : 100 94 90 96 91 94 95 99 95 98 100 97 99 98 96 98 ... Зададим n = 4, m = 1, s = 1. С помощью метода окон для нейронной сети будет сгенерирована следующая обучающая выборка: 100 94 90 96 -> 91 94 90 96 91 -> 94 90 96 91 94 -> 95 96 91 94 95 -> 99 91 94 95 99 -> 95 и т.д.

Каждый следующий вектор получается в результате сдвига окон W_i и W_o вправо на один элемент (s = 1) . Предполагается наличие скрытых зависимостей во временной последовательности как множестве наблюдений. Нейронная сеть, обучаясь на этих наблюдениях и соответственно настраивая свои коэффициенты, пытается извлечь эти закономерности и сформировать в результате требуемую функцию прогноза P.

Прогнозирование осуществляется по тому же принципу, что и формирование обучающей выборки. При этом выделяются две возможности: одношаговое и многошаговое прогнозирование.

Многошаговое прогнозирование

Используется для осуществления долгосрочного прогноза и предназначено для определения основного тренда и главных точек изменения тренда для некоторого промежутка времени в будущем. При этом прогнозирующая система использует полученные (выходные) данные для моментов времени k+1, k+2 и т.д. в качестве входных данных для прогнозирования на моменты времени k+2, k+3 и т.д.

Предположим, система обучилась на временной последовательности. Затем она спрогнозировала k+1 элемент последовательности, например, равный 95, когда на ее вход был подан последний из известных ей образов (99,98,96,98) . После этого она осуществляет дальнейшее прогнозирование и на вход подается следующий образ (98,96,98,95) . Последний элемент этого образа является прогнозом системы. И так далее.

Одношаговое прогнозирование

Используется для краткосрочных прогнозов, обычно - абсолютных значений последовательности. Осуществляется прогноз только на один шаг вперед, но используется реальное, а не прогнозируемое значение для осуществления прогноза на следующем шаге.

Для временной последовательности 3.2. На шаге k+1 система прогнозирует требование 95, хотя реальное значение (смотри приложение 1) должно быть 96. На шаге k + 2 в качестве входного образа будет использоваться образ (98,96,98,96) .

Как было сказано выше, результатом прогноза на НС является класс к которому принадлежит переменная, а не ее конкретное значение. Формирование классов должно проводиться в зависимости от того каковы цели прогнозирования. Общий подход состоит в том, что область определения прогнозируемой переменной разбивается на классы в соответствии с необходимой точностью прогнозирования. Классы могут представлять качественный или численный взгляд на изменение переменной.

Вопросы по лекции

1. Дайте описание принципиальной концепции построения нейронных сетей.
2. Какие типы нейросетевых структур используются в прогнозировании?
3. Опишите классификацию нейронных сетей.
4. Возможности и области применения персептронов.
5. Процедура подбора, предобработки обучающих примеров и их интерпретация.
6. Понятие, назначение и виды активационных функций.
7. Особенности нейросетевого прогнозирования как алгоритма.
8. Обоснование целесообразности нейросетевого прогнозирования .
9. Сущность генетических методов прогнозирования.
10. Принципы прогнозирования на основе самоорганизующих крат Кохонена.

Тема 2.2 Основы имитационного моделирования

Лекция №7 Моделирование сложных систем

1. Понятие системы, класса и экземпляра.
2. Генетические алгоритмы в моделировании систем
3. Процессный подход к моделированию
4. Моделирование систем массового обслуживания
5. Сети Петри

1. Понятие системы, класса и экземпляра.

Под системой будем понимать целостную совокупность взаимосвязанных компонент (объектов), которая характеризуется организованностью, общим поведением и устойчивостью. Организованность системы проявляется в связях и взаимодействиях составляющих ее подсистем и объектов. Целостность позволяет рассматривать систему как некую единую сущность, как единый объект при анализе ее взаимодействий со средой. Устойчивость обеспечивает идентификацию системы в процессе ее эволюции. Таким образом, система – это устойчивая во времени и ограниченная в пространстве совокупность объектов, интересующая нас как единое целое

Система определяется структурой и поведением. Эти понятия для системологии являются такими же фундаментальными, изначальными, неопределяемыми понятиями как для физики пространство и время. Под структурой понимается инвариантная во времени фиксация связей между элементами системы (статическое описание системы).

Под поведением системы понимается ее функционирование во времени (динамическое описание системы). Поведенческие аспекты системы выражаются с помощью понятий функция и цель. В самом общем плане цель – это состояние (совокупность свойств системы, определяющих характер её отношений с внешней средой), к которому направлена тенденция движения системы. Заметим, что с усложнением иерархического строения системы возрастает и сложность ее поведения. Это связано с возрастанием сложности неабстрактной структуры систем.

Статическое или структурное описание – это совокупность утверждений, которые связывают значения фазовых координат системы друг с другом при определенном выборе фазового пространства системы StΣ.

Динамическое описание, напротив, указывает, как изменение значений фазовых координат зависит от значений или изменений значений других координат. Таким образом, динамическое описание – это совокупность утверждений, из которых можно математически вывести поведение системы: описать, как она переходит из одного состояния в другое.

Динамическое описание рассматривает систему Σ как структуру, на которую в определенные моменты t $\in$ T оказывается входное воздействие x(t) (например, поступает вещество, энергия или информация) и которая в какие-то моменты времени порождает некоторую выходную величину y(t). Предполагается, что значения входного воздействия x(t) $\in$ X выбираются из некоторого фиксированного множества X (то есть в любой момент времени x(t) должно принадлежать множеству X), а любое мгновенное значение выходной величины y(t) – принадлежать некоторому фиксированному множеству Y.

Причем, в общем случае, значение выходной величины системы Σ зависит как от текущего входного воздействия, так и от предыстории этого воздействия, которое сказывается на внутреннем состоянии q(t) $\in$ Q. Другими словами, внутреннее состояние системы q(t) $\in$ Q, определяет текущее значение выходной величины системы y(t) $\in$Y и оказывает влияние на ее будущее. Таким образом, при таком подходе внутреннее состояние системы q(t) рассматривается как память, или накопитель прецедентов системы Σ, а математическое описание динамической системы является описанием потока причинно-следственных связей из прошлого в будущее.

Понятие класса как множества сущностей (объектов), имеющих одинаковое функциональное назначение, структуру и поведение и отличающихся значением параметров, является ключевым для проектирования многокомпонентных систем. С понятием класса неразрывно связано понятие экземпляра класса, т. е. конкретного объекта из множества всех объектов того же самого класса с уникальными значениями параметров. Разбиение реальной или проектируемой системы на компоненты, построение классов, соответствующих этим компонентам, установление связей между компонентами, аналогичных связям между компонентами моделируемой реальной или проектируемой системы, и построение модели из экземпляров классов с учетом существующих связей — это ключевые моменты технологии объектно-ориентированного моделирования. При объектно-ориентированного моделирования. При объектно-ориентированном подходе структура модели соответствует структуре реального объекта.

Использование экземпляров класса позволяет собирать новую модель из типовых компонентов, подстраивая их каждый раз под конкретные условия функционирования, а также использовать построенную модель в качестве компонента других моделей.

2. Генетические алгоритмы в моделировании систем

Генетические алгоритмы являются частью более общей группы методов, называемой эволюционными вычислениями, которые объединяют различные варианты использования эволюционных принципов для достижения поставленной цели.

Также в ней выделяют следующие направления:

- Эволюционные стратегии

- Метод оптимизации, основанный на идеях адаптации и эволюции. Степень мутации в данном случае меняется со временем – это приводит к, так называемой, самоадаптации.
- Генетическое программирование
- Применение эволюционного подхода к популяции программ.
- Эволюционное программирование

Было впервые предложено Л.Дж. Фогелем в 1960 году для моделирования эволюции как процесса обучения с целью создания искусственного интеллекта. Он использовал конечные автоматы, предсказывающие символы в цифровых последовательностях, которые, эволюционируя, становились более приспособленными к решению поставленной задачи.

Генетические алгоритмы применяются для решения следующих задач:

- Оптимизация функций
- Разнообразные задачи на графах (задача коммивояжера, раскраска, нахождение паросочетаний)
- Настройка и обучение искусственной нейронной сети
- Задачи компоновки
- Составление расписаний
- Игровые стратегии
- Аппроксимация функций
- Искусственная жизнь
- Биоинформатика

Классический ГА

Постановка задачи и функция приспособленности

Пусть перед нами стоит задача оптимизации, например:

I. Задача наилучшего приближения

Если рассматривать систему n линейных уравнений с m неизвестными

$$Ax = b$$

в случае, когда она переопределена ($n > m$), то иногда оказывается естественной задача о нахождении вектора x, который "удовлетворяет этой системе наилучшим образом", т. е. из всех "не решений" является лучшим.

II. Задача о рационе.

Пусть имеется n различных пищевых продуктов, содержащих m различных питательных веществ. Обозначим через aij содержание (долю) j-го питательного вещества в i-ом продукте, через bj — суточную потребность организма в j-ом питательном веществе, через ci — стоимость единицы i-го продукта. Требуется составить суточный рацион питания минимальной стоимости, удовлетворяющий потребность во всех питательных веществах

III. Транспортная задача.

Эта задача — классическая задача линейного программирования. К ней сводятся многие оптимизационные задачи. Формулируется она так. На m складах находится груз, который нужно развезти n потребителям. Пусть ai ($i = 1, ..., n$) — количество груза на i-ом складе, а bj ($j = 1, ..., m$) — потребность в

грузе j-го потребителя, cij — стоимость перевозки единицы груза с i-го склада j-му потребителю. Требуется минимизировать стоимость перевозок.

IV. Задачи о распределении ресурсов.

Общий смысл таких задач — распределить ограниченный ресурс между потребителями оптимальным образом. Рассмотрим простейший пример — задачу о режиме работы энергосистемы. Пусть m электростанций питают одну нагрузку мощности p. Обозначим через xj активную мощность, генерируемую j-ой электростанцией. Техническими условиями определяются возможный минимум mj и максимум Mj вырабатываемой j-ой электростанцией мощности. Допустим затраты на генерацию мощности x на j-ой электростанции равны ej(x). Требуется сгенерировать требуемую мощность p при минимальных затратах.

Переформулируем задачу оптимизации как задачу нахождения максимума некоторой функции f(x1, x2, …, xn), называемой функцией приспособленности (fitness function). Она должна принимать неотрицательные значения на ограниченной области определения (для того, чтобы мы могли для каждой особи считать её приспособленность, которая не может быть отрицательной), при этом совершенно не требуются непрерывность и дифференцируемость.

Каждый параметр функции приспособленности кодируется строкой битов.

Особью будет называться строка, являющаяся конкатенацией строк упорядоченного набора параметров:

1010 10110 101 … 10101

| x1 | x2 | x3 | … | xn |

Универсальность ГА заключается в том, что от конкретной задачи зависят только такие параметры, как функция приспособленности и кодирование решений. Остальные шаги для всех задач производятся одинаково.

Принцип работы ГА

Генетические алгоритмы оперируют совокупностью особей (популяцией), которые представляют собой строки, кодирующие одно из решений задачи. Этим ГА отличается от большинства других алгоритмов оптимизации, которые оперируют лишь с одним решением, улучшая его.

С помощью функции приспособленности среди всех особей популяции выделяют:

наиболее приспособленные (более подходящие решения), которые получают возможность скрещиваться и давать потомство

наихудшие (плохие решения), которые удаляются из популяции и не дают потомства

Таким образом, приспособленность нового поколения в среднем выше предыдущего.

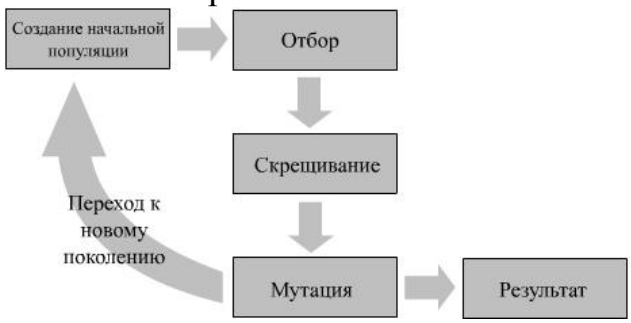
В классическом ГА:

- начальная популяция формируется случайным образом
- размер популяции (количество особей N) фиксируется и не изменяется в течение работы всего алгоритма

- каждая особь генерируется как случайная L-битная строка, где L — длина кодировки особи
- длина кодировки для всех особей одинакова

Алгоритм работы

На рисунке изображена схема работы любого генетического алгоритма:



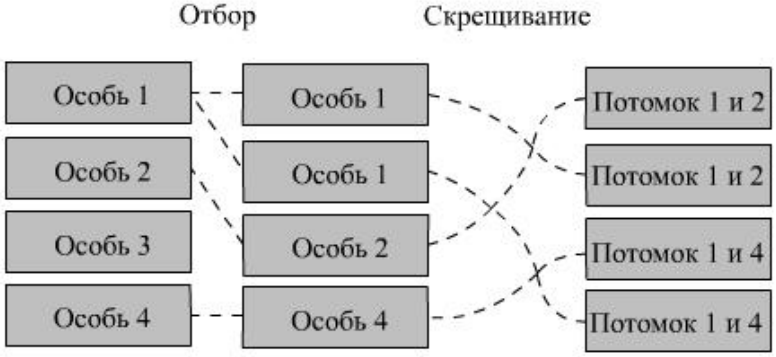
Шаг алгоритма состоит из трех стадий:

генерация промежуточной популяции (intermediate generation) путем отбора (selection) текущего поколения

скрещивание (recombination) особей промежуточной популяции путем кроссовера (crossover), что приводит к формированию нового поколения

мутация нового поколения

Первые две стадии (отбор и скрещивание):



Отбор

Промежуточная популяция — это набор особей, получивших право размножаться. Наиболее приспособленные особи могут быть записаны туда несколько раз, наименее приспособленные с большой вероятностью туда вообще не попадут.

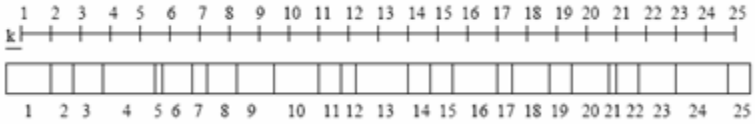
В классическом ГА вероятность каждой особи попасть в промежуточную популяцию пропорциональна ее приспособленности, т.е. работает пропорциональный отбор (proportional selection).

Существует несколько способов реализации данного отбора:

stochastic sampling. Пусть особи располагаются на колесе рулетки так, что размер сектора каждой особи пропорционален ее приспособленности. N раз запуская рулетку, выбираем требуемое количество особей для записи в промежуточную популяцию.

remainder stochastic sampling. Для каждой особи вычисляется отношение ее приспособленности к средней приспособленности популяции. Целая часть

этого отношения указывает, сколько раз нужно записать особь в промежуточную популяцию, а дробная показывает её вероятность попасть туда ещё раз. Реализовать такой способ отбора удобно следующим образом: расположим особи на рулетке так же, как было описано. Теперь пусть у рулетки не одна стрелка, а N, причем они отсекают одинаковые сектора. Тогда один запуск рулетки выберет сразу все N особей, которые нужно записать в промежуточную популяцию. Такой способ иллюстрируется следующим рисунком:



Скрещивание

Особи промежуточной популяции случайным образом разбиваются на пары, потом с некоторой вероятностью скрещиваются, в результате чего получаются два потомка, которые записываются в новое поколение, или не скрещиваются, тогда в новое поколение записывается сама пара.

В классическом ГА применяется одноточечный оператор кроссовера (1-point crossover): для родительских строк случайным образом выбирается точка раздела, потомки получаются путём обмена отсечёнными частями.

011010.01010001101 -> 111100.01010001101

111100.10011101001 011010.10011101001

Мутация

К полученному в результате отбора и скрещивания новому поколению применяется оператор мутации, необходимый для "выбивания" популяции из локального экстремума и способствующий защите от преждевременной сходимости.

Каждый бит каждой особи популяции с некоторой вероятностью инвертируется. Эта вероятность обычно очень мала, менее 1%.

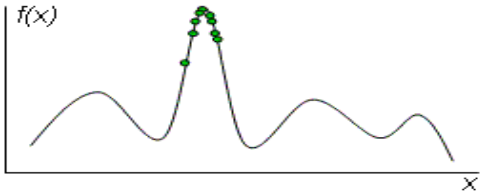
1011001100101101 -> 1011001101101101

Можно выбирать некоторое количество точек в хромосоме для инверсии, причем их число также может быть случайным. Также можно инвертировать сразу некоторую группу подряд идущих точек. Среди рекомендаций по выбору вероятности мутации нередко можно встретить варианты 1/L или 1/N.

Критерии останова

Такой процесс эволюции, вообще говоря, может продолжаться до бесконечности. Критерием останова может служить заданное количество поколений или схождение (convergence) популяции.

Схождением называется состояние популяции, когда все строки популяции находятся в области некоторого экстремума и почти одинаковы. То есть кроссовер практически никак не изменяет популяции, а мутирующие особи склонны вымирать, так как менее приспособлены. Таким образом, схождение популяции означает, что достигнуто решение близкое к оптимальному.



Итоговым решением задачи может служить наиболее приспособленная особь последнего поколения.

Генетический алгоритм производит поиск решений с помощью:

- отбора гиперплоскостей (hyperplane sampling), осуществляемого кроссовером, поскольку последний комбинирует и совмещает шаблоны родителей в их детях.
- метода hill-climbing, обеспечивающегося мутацией: особь случайным образом изменяется – неудачные варианты вымирают, полезные изменения сохраняются популяцией.

Исследования показали, что на простых задачах с малым размером популяции ГА с мутацией (и без кроссовера) находят решение быстрее. На сложных многоэкстремальных функциях лучше использовать ГА с кроссовером, поскольку этот метод более надежен, хотя и требует большего размера популяции.

С точки зрения теоремы шаблонов, мутация только вредит росту количества представителей хороших шаблонов, поскольку лишний раз их разрушает. Однако мутация просто необходима для ГА с малым размером популяции, потому что для них свойственна преждевременная сходимость (premature convergence) – ситуация, когда в некоторых позициях все особи имеют один и тот же бит, не соответствующий глобальному экстремуму.

Введём понятие, использующееся для анализа ГА. Давление отбора (selection pressure) — это мера того, насколько различаются шансы лучшей и худшей особей популяции попасть в промежуточную популяцию. Для пропорционального отбора эта величина с увеличением средней приспособленности популяции уменьшается, стремясь к 1, т.е. шансы плохой и хорошей особей создать потомство уравниваются.

При увеличении вероятностей скрещивания или мутации и при уменьшении давления отбора (например, за счет использования других стратегий отбора) размножение представителей приспособленных шаблонов замедляется, но зато происходит интенсивный поиск других шаблонов. Обратно, уменьшение вероятностей скрещивания или мутации и увеличение давления отбора ведет к интенсивному использованию найденных хороших шаблонов, но меньше внимания уделяется поиску новых.

Вывод: для эффективной работы генетического алгоритма необходимо поддерживать тонкое равновесие между исследованием и использованием.

3. Процессный подход к моделированию

Процессный подход предполагает разбиение объекта на отдельные процессы и их дальнейшая декомпозиция с учетом факторов, механизмов и результатов каждого отдельного процесса. Методология моделирования может быть рассмотрена на примере классификации моделей в таблице 7.1.

Таблица 7.1 - Классификация процессных моделей

Модель	Характеристика	Область моделирования
SADT	совокупность методов, правил и процедур, предназначенных для построения функциональной модели объекта какой-либо предметной области. Функциональная модель SADT отображает функциональную структуру объекта, т.е. производимые им действия и связи между этими действиями.	Методология SADT может использоваться для моделирования широкого круга систем и определения требований и функций, а затем для разработки системы, которая удовлетворяет этим требованиям и реализует эти функции. Для уже существующих систем, SADT может быть использована для анализа функций, выполняемых системой, а также для указания механизмов, посредством которых они осуществляются.
IDEF0	используется для создания функциональной модели, отображающей структуру и функции системы, а также потоки информации и материальных объектов, связывающие эти функции;	предназначена для описания существующих бизнес-процессов, в котором используются как естественный, так и графический языки. Для передачи информации о конкретной системе источником графического языка является сама методология IDEFO
DFD	методология предназначенная для описания потоков данных. Они позволяют отразить последовательность работ, выполняемых по ходу процесса, и и потоки информации, циркулирующие между этими работами. Кроме того, нотация DFD предоставляет возможность описывать потоки документов (документооборот) и материальных ресурсов (например, движение материалов от одной работы к другой)	Методология DFD может эффективно использоваться для описания процессов при внедрении процессного подхода к управлению организацией, так как позволяет максимально снизить субъективность описания бизнес-процессов. С помощью схемы процессов в DFD выявляют основные потоки данных, что важно для последующего создания моделей структуры данных и разработки требований к информационной системе организации
IDEF3	методология документирования процессов, происходящих в системе, которая используется, например, при исследовании технологических процессов на предприятиях. IDEF3, предназначенная для описания рабочих процессов или, иными словами, потоков работ (Work Flow Modeling)	С помощью IDEF3 описываются сценарий и последовательность операций для каждого процесса. IDEF3 имеет прямую взаимосвязь с методологией IDEF0 – каждая функция может быть представлена в виде отдельного процесса средствами IDEF3. Можно утверждать, что IDEF3 лежит в основе популярной в настоящее время методологии ARIS eEPC
IDEFI	применяется для построения информационной модели,	Метод IDEFI основан на подходе Чена и позволяет построить модель данных,

	отображающей структуру и содержание информационных потоков, необходимых для поддержки функций системы	эквивалентную реляционной модели в третьей нормальной форме. На основе совершенствования метода IDEFI создана его новая версия — метод IDEFIX, разработанный с учетом таких требований, как простота для изучения и возможность автоматизации. IDEFIX-диаграммы используются в ряде распространенных CASE-средств (в частности, ERwin, Design/IDEF).

1) Методология SADT представляет собой совокупность методов, правил и процедур, предназначенных для построения функциональной модели объекта какой-либо предметной области. Функциональная модель SADT отображает функциональную структуру объекта, т.е. производимые им действия и связи между этими действиями.

Результатом применения методологии SADT является модель, которая состоит из диаграмм, фрагментов текстов и глоссария, имеющих ссылки друг на друга. Диаграммы - главные компоненты модели, все функции ИС и интерфейсы на них представлены как блоки и дуги. Место соединения дуги с блоком определяет тип интерфейса. Управляющая информация входит в блок сверху, в то время как информация, которая подвергается обработке, показана с левой стороны блока, а результаты выхода показаны с правой стороны. Механизм (человек или автоматизированная система), который осуществляет операцию, представляется дугой, входящей в блок снизу (рис.7.1).

Одной из наиболее важных особенностей методологии SADT является постепенное введение все больших уровней детализации по мере создания диаграмм, отображающих модель.

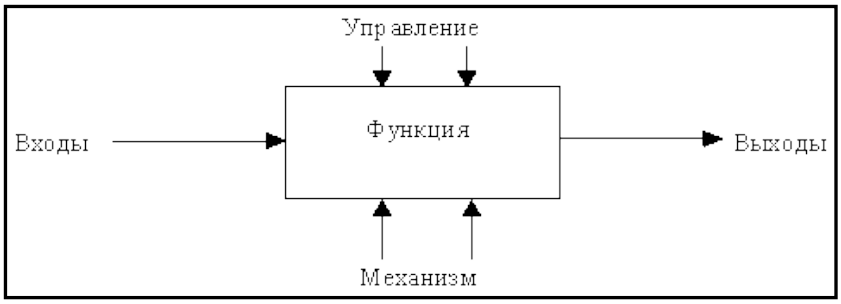


Рисунок 7.1 – Модель SADT

2) На основе методология SADT разработана, методология IDEF0, которая была утверждена в качестве федерального стандарта США в 1993 г.

Методологию IDEF0 можно считать следующим этапом развития хорошо известного графического языка описания функциональных систем SADT (Structured Analysis and Design Technique). Исторически IDEF0 как стандарт был разработан в 1981 году в рамках обширной программы, когда ВВС США предложили и реализовали Программу интегрированной компьютеризации производства ICAM (ICAM - Integrated Computer Aided Manufacturing),направленную на увеличение эффективности промышленных предприятий посредством широкого внедрения компьютерных (информационных) технологий.

С помощью методологии семейства IDEF можно эффективно отображать и анализировать модели деятельности широкого спектра сложных систем в различных разрезах. При этом широта и глубина обследования процессов в системе определяется самим разработчиком, что позволяет не перегружать создаваемую модель излишними данными. В настоящий момент к семейству IDEF можно отнести следующие стандарты:

1.IDEF0 используется для создания функциональной модели, отображающей структуру и функции системы, а также потоки информации и материальных объектов, связывающие эти функции;

2.IDEF1 применяется для построения информационной модели, отображающей структуру и содержание информационных потоков, необходимых для поддержки функций системы;

3.IDEF1X (IDEF1 Extended) – методология построения реляционных структур. IDEF1X относится к типу методологий “Сущность-взаимосвязь” (ER – Entity-Relationship) и, как правило, используется для моделирования реляционных баз данных;

4.IDEF2 позволяет построить динамическую модель меняющихся во времени поведения функций, информации и ресурсов системы.

5.IDEF3 – методология документирования процессов, происходящих в системе, которая используется, например, при исследовании технологических процессов на предприятиях. IDEF3, предназначенная для описания рабочих процессов или, иными словами, потоков работ (Work Flow Modeling). С помощью IDEF3 описываются сценарий и последовательность операций для каждого процесса. IDEF3 имеет прямую взаимосвязь с методологией IDEF0 – каждая функция может быть представлена в виде отдельного процесса средствами IDEF3. Можно утверждать, что IDEF3 лежит в основе популярной в настоящее время методологии ARIS eEPC.

6.DEF4 – методология построения объектно-ориентированных систем. Средства IDEF4 позволяют наглядно отображать структуру объектов и заложенные принципы их взаимодействия, тем самым позволяя анализировать и оптимизировать сложные объектно-ориентированные системы;

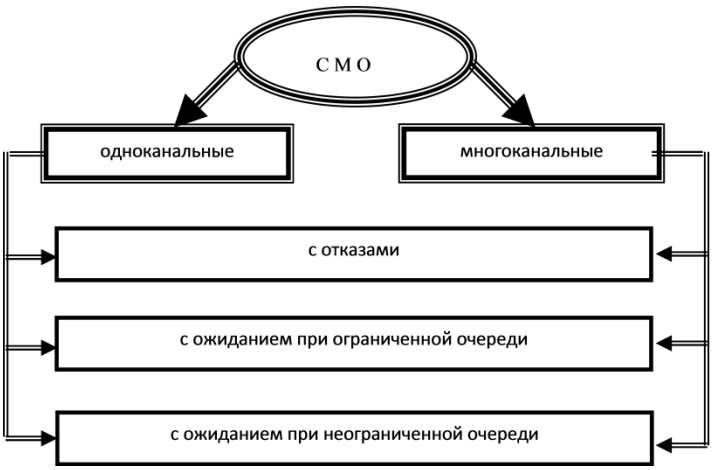
7.IDEF5 – методология исследования сложных систем.

8.DFD (Data Flow Diagramming), методология предназначенная для описания потоков данных. Они позволяют отразить последовательность работ, выполняемых по ходу процесса, и потоки информации, циркулирующие между этими работами.

4. Моделирование систем массового обслуживания

Системы, в которых в случайные моменты времени возникают заявки на обслуживание и имеются устройства для обслуживания этих заявок, называются системами массового обслуживания (СМО).

СМО могут быть классифицированы по признаку организации обслуживания следующим образом:



Системы с отказами не имеют очередей.

Системы с ожиданием имеют очереди.

Заявка, поступившая в момент, когда все каналы обслуживания заняты:

- покидает систему с отказами;
- становится в очередь на обслуживание в системах с ожиданием при неограниченной очереди или на свободное место при ограниченной очереди;
- покидает систему с ожиданием при ограниченной очереди, если в этой очереди нет свободного места.

В качестве меры эффективности экономической СМО рассматривают сумму потерь времени:

- на ожидание в очереди;
- на простои каналов обслуживания.

Для всех видов СМО используются следующие показатели эффективности:

- относительная пропускная способность - это средняя доля поступающих заявок, обслуживаемых системой;
- абсолютная пропускная способность - это среднее число заявок, обслуживаемых системой в единицу времени;
- вероятность отказа - это вероятность того, что заявка покинет систему без обслуживания;
- среднее число занятых каналов - для многоканальных СМО.

Показатели эффективности СМО рассчитываются по формулам из специальных справочников (таблиц). Исходными данными для таких расчетов являются результаты моделирования СМО.

Моделирование системы массового обслуживания: основные параметры, граф состояний

При всем многообразии СМО они имеют общие черты, которые позволяют унифицировать их моделирование для нахождения наиболее эффективных вариантов организации таких систем.

Для моделирования СМО необходимо иметь следующие исходные данные:

- основные параметры;
- граф состояний.

Результатами моделирования СМО являются вероятности ее состояний, через которые выражаются все показатели ее эффективности.

Основные параметры для моделирования СМО включают:

- характеристики входящего потока заявок на обслуживание;
- характеристики механизма обслуживания.

Рассмотрим характеристики потока заявок.

Поток заявок - последовательность заявок, поступающих на обслуживание.

Интенсивность потока заявок λ - среднее число заявок, поступающих в СМО в единицу времени.

Потоки заявок бывают простейшими и отличными от простейших.

Для простейших потоков заявок используются модели СМО.

Простейшим, или пуассоновским называется поток, являющийся стационарным, одинарным и в нем отсутствуют последействия.

Стационарность означает неизменность интенсивности поступления заявок с течением времени.

Одинарным поток заявок является в том случае, когда за малый промежуток времени вероятность поступления более чем одной заявки близка к нулю.

Отсутствие последействия заключается в том, что число заявок, поступивших в СМО за один интервал времени, не влияет на количество заявок, полученных за другой интервал времени.

Для отличных от простейших потоков заявок используются имитационные модели.

Рассмотрим характеристики механизма обслуживания.

Механизм обслуживания характеризуется:

- числом n каналов обслуживания;
- производительностью канала, или интенсивностью обслуживания μ - средним числом заявок, обслуживаемых одним каналом в единицу времени;
- дисциплиной очереди (например, объемом очереди m, порядком отбора из очереди в механизм обслуживания и т.п.).

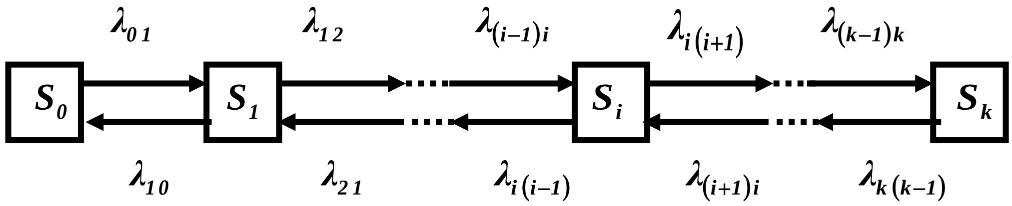
Граф состояний описывает функционирование системы обслуживания как переходы из одного состояния в другое под действием потока заявок и их обслуживания.

Для построения графа состояний СМО необходимо:

- составить перечень всех возможных состояний СМО;
- представить перечисленные состояния графически и отобразить возможные переходы между ними стрелками;
- взвесить отображенные стрелки, т.е. приписать им числовые значения интенсивностей переходов, определяемые интенсивностью потока заявок и интенсивностью их обслуживания.

Вычисление вероятностей состояний системы массового обслуживания

Граф состояний СМО со схемой "гибели и рождения" представляет собой линейную цепочку, где каждое из средних состояний имеет прямую и обратную связь с каждым из соседних состояний, а крайние состояния только с одним соседним:



Число состояний в графе на единицу больше, чем суммарное число каналов обслуживания и мест в очереди.

СМО может быть в любом из своих возможных состояний, поэтому ожидаемая интенсивность выхода из какого-либо состояния равна ожидаемой интенсивности входа системы в это состояние. Отсюда система уравнений для определения вероятностей состояний при простейших потоках будет иметь вид:

$$\begin{cases} \lambda_{0\,1} \cdot P_0 = \lambda_{1\,0} \cdot P_1, \\ \left(\lambda_{1\,0} + \lambda_{1\,2}\right) \cdot P_1 = \lambda_{0\,1} \cdot P_0 + \lambda_{2\,1} \cdot P_2, \\ \cdots\cdots\cdots\cdots\cdots\cdots , \\ \left(\lambda_{i\,(i-1)} + \lambda_{i\,(i+1)}\right) \cdot P_i = \lambda_{(i-1)\,i} \cdot P_{i-1} + \lambda_{(i+1)\,i} \cdot P_{i+1}, \\ \cdots\cdots\cdots\cdots\cdots\cdots , \\ \left(\lambda_{(k-1)\,(k-2)} + \lambda_{(k-1)\,k}\right) \cdot P_{k-1} = \lambda_{(k-2)\,(k-1)} \cdot P_{k-2} + \lambda_{k\,(k-1)} \cdot P_k, \\ \lambda_{k\,(k-1)} \cdot P_k = \lambda_{(k-1)\,k} \cdot P_{k-1}; \end{cases}$$

где P_i - вероятность того, что система находится в состоянии S_i, $i = \overline{0,k}$;

$\lambda_{i(i+1)}$ $\left(\lambda_{i(i-1)}\right)$ - интенсивность перехода, или среднее число переходов системы в единицу времени из состояния S_i в состояние S_{i+1} $\left(S_{i-1}\right)$.

Используя эту систему уравнений, а также уравнение

$$\sum_{i=0}^{k} P_i = 1,$$

вероятность P_i любого i -ого состояния $\left(i = \overline{0,k}\right)$ можно вычислить по следующему общему правилу:

вероятность нулевого состояния рассчитывается как

$$P_0 = \left(1 + \lambda_{01}/\lambda_{10} + \lambda_{01} \cdot \lambda_{12}/\lambda_{21} \cdot \lambda_{10} + \ldots + \lambda_{01} \cdot \lambda_{12} \cdots \lambda_{(i-1)i}/\lambda_{i(i-1)} \cdots \lambda_{21} \cdot \lambda_{10} + \right.$$
$$\left. + \ldots + \lambda_{01} \cdot \lambda_{12} \cdots \lambda_{(i-1)i} \cdots \lambda_{(k-1)k}/\lambda_{k(k-1)} \cdots \lambda_{i(i-1)} \cdots \lambda_{21} \cdot \lambda_{10}\right)^{-1},$$

а затем берется дробь, в числителе которой стоит произведение всех интенсивностей потоков по стрелкам, ведущим слева направо от состояния S_0 до состояния S_i, а в знаменателе - произведение всех интенсивностей по

стрелкам, идущим справа налево от состояния S_i до состояния S_0, и эта дробь умножается на рассчитанную вероятность P_0

$$P_i = \left(\lambda_{01} \cdot \lambda_{12} \cdots \lambda_{(i-1)i} \,/\, \lambda_{i(i-1)} \cdots \lambda_{21} \cdot \lambda_{10}\right) \cdot P_0 .$$

Системы массового обслуживания имеют один или несколько каналов обслуживания и могут иметь ограниченную или неограниченную очередь (системы с ожиданием) заявок на обслуживание, не иметь очереди (системы с отказами). Заявки на обслуживание возникают в случайные моменты времени. Системы массового обслуживания характеризуются следующими показателями эффективности: относительная пропускная способность, абсолютная пропускная способность, вероятность отказа, среднее число занятых каналов.

Моделирование систем массового обслуживания осуществляется для нахождения наиболее эффективных вариантов их организации и предполагает следующие исходные данные для этого: основные параметры, граф состояний. К таким данным относятся следующие: интенсивность потока заявок, количество каналов обслуживания, интенсивность обслуживания и объем очереди. Число состояний в графе на единицу больше, чем сумма числа каналов обслуживания и мест в очереди.

Вычисление вероятностей состояний системы массового обслуживания со схемой «гибели и рождения» осуществляется по общему правилу.

5. Сети Петри

Сети Петри используются для моделирования асинхронных систем, функционирующих как совокупность параллельных взаимодействующих процессов. Анализ сетей Петри позволяет получить информацию о структуре и динамическом поведении моделируемой системы.

Причинно-следственная связь событий в асинхронных системах задается множеством отношений вида "условия-события".

Построение моделей систем в виде сетей Петри заключается в следующем:

1. Моделируемые процессы описываются множеством событий (действий) и условий определяющих возможность наступления этих событий, а также причинно-следственными отношениями, устанавливаемыми на множестве пар "события-условия".

2. Определяются события-действия, последовательность выполнения которых управляется состояниями системы. Состояния системы задаются множеством условий, формируемых в виде предикатов. Количественно условия характеризуются величиной которая выражается числами натурального ряда.

3. Условия, в зависимости от значений их количественных характеристик, могут выполняться или нет. Выполнение условий обеспечивает возможность реализации событий. Условия, с фактом выполнения которых связывается возможность реализации событий, называются предусловиями. Реализация события обеспечивает возможность выполнения других условий, находящихся

с предусловиями в причинно-следственной связи. Эти условия называются постусловиями.

В сетях Петри условия - это позиции, а события - переходы. В соответствии с этим граф сети Петри является двудольным ориентированным мультиграфом. Изображение позиции и перехода на графе показано на рисунке 5.1.

а) б)

Рис 7.2. а) − изображение позиции; б) − изображение перехода.

Ориентированные дуги могут соединять только позиции и переходы в прямом и обратном направлении (свойство двудольности). Сеть Петри является мультиграфом, так как допускается кратность дуг между позициями и переходами (вершинами графа). Пример графа сети Петри приведен на рис.5.2.

В графах сети Петри количественные характеристики условий (числа натурального ряда) принято изображать числом меток в соответствующих позициях (см. рис.7.3).

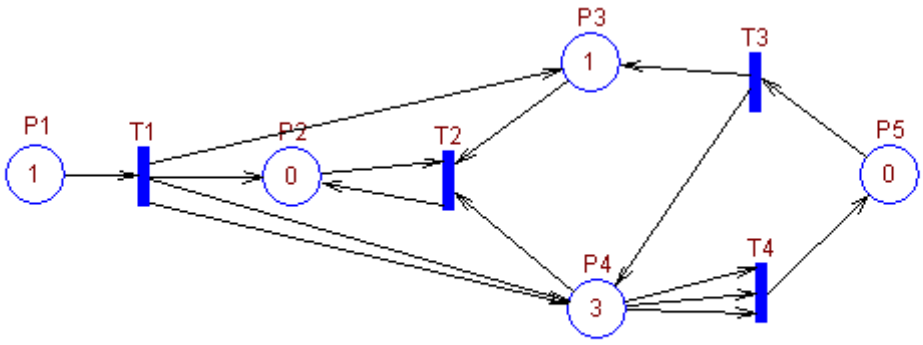


Рис.7.3. Пример графа сети Петри.

Последовательности событий отображаются срабатываниями переходов. Выполнение какого-либо условия связано с появлением одной или нескольких меток в соответствующей этому условию позиции. Соглашения о правилах срабатывания переходов является способом выражения причинно-следственных связей между условиями и событиями в системе.

Формализованное описание сетей Петри

Для описания и математического анализа процессов с точками ветвления и синхронизации взаимодействия разработан аппарат сетей Петри. Структура сети представляется ориентированным двудольным графом. Множество V вершин графа разбивается на два подмножества T и P, $V = T \cup P$, $T \cap P = \varnothing$. Дугами могут связываться вершины из множеств P и T. Динамика развития процессов отражается в вершинах P метками (марками). Распределение меток по вершинам P называют маркированием сети. Каждое маркирование соответствует определенному состоянию сети.

Сеть Петри определяется пятеркой $N = \{P, T, J, O, M_0\}$, где

$P = \{p_i\}$, $i = 1,2,...,n$ - множество позиций;

$T = \{t_j\}$, $j = 1,2,...,m$ - множество переходов;

$J : T \times P \to \{0,1\}$ - функция следования;

$O : P \times T \to \{0,1\}$ - функция предшествования;

$M_0 : P \to Z_0$ - начальное маркирование (состояние) сети;

Z_0 - множество положительных целых чисел.

Функции J и O задают множества дуг $\{t_j, p_i\}$ и $\{p_i, t_j\}$ соответственно.

Дуги, предшествующие позиции p_i, обозначим множеством $J(p_i) = \{(t_j, p_i) | J(t_j, p_i) = 1\}$, а дуги, предшествующие переходу t_j, множеством $J(t_j) = \{(p_i, t_j) | O(p_i, t_j) = 1\}$.

Здесь запись $J(t_j, p_i) = 1$ означает наличие дуги (t_j, p_i), а запись $O(p_i, t_j) = 1$ - дуги (p_i, t_j). Аналогично, дуги, следующие из p_i и t_j, представим множествами $O(p_i) = \{(p_i, t_j) | O(p_i, t_j) = 1\}$, $O(t_j) = \{(t_j, p_i) | J(t_j, p_i) = 1\}$.

Входные позиции перехода t_j объединяются в множества его предшественников $\mathrm{Pr}e(t_j) = \{p_i \in P | O(p_i, t_j) = 1\}$, а выходные позиции – в множества позиций–последователей $Post(t_j) = \{p_i \in P | J(t_j, p_i) = 1\}$.

Маркирование сети представляется вектором $M = \{m(p_i)\}$, где $m(p_i)$ - число меток в позиции p_i. Переход t_j возбужден при маркировании M и может сработать, если выполняется условие $\forall p_i \in \mathrm{Pr}e(y_j) [m(p_i) - O(p_i, t_j) \geq 0]$, то есть число меток $m(p_i)$ больше или равно числу дуг (p_i, t_j), что соответствует $m(p_i) - O(p_i, t_j) \geq 0$.

Срабатывание перехода t_j приводит к тому, что каждая позиция $p_i \in \mathrm{Pr}e(t_j)$ теряет $O(p_i, t_j)$ меток, а каждая из позиций $Post(t_j)$ получает $J(t_j, p_i)$ меток.

Если при маркировании M возбуждено несколько переходов, то порядок их срабатывания не определен, и, следовательно, может быть представлено несколько последовательностей срабатывающих переходов.

Вопросы по лекции

1. Дайте понятие и опишите свойства системы.
2. Понятие классов, экземпляров и многокомпонентных систем.
3. Понятие и сущность генетических алгоритмов.
4. Принципы создания новых компонентов модели.
5. Описание SADT-методологии проектирования сложных систем.
6. Функциональное моделирование в нотации IDEF0.
7. Опишите методологию событийного моделирования IDEF3.

8. Опишите методологию моделирования потоков данных.
9. Описание классификации систем массового обслуживания.
10. Набор и основные элементы сети Петри.

Лекция № 8 Линейное и нелинейное программирование

План лекции

1. Общий вид математической модели задачи линейного программирования
2. Математическая формулировка двойственной задачи к произвольной задаче линейного программирования
3. Нелинейное программирование

1. Общий вид математической модели задачи линейного программирования

В общем виде задача линейного программирования ставится следующим образом:

Найти набор управляемых параметров

$$\bar{x} = (x_1, x_2, ...x_n),$$

на котором достигается наибольшее (наименьшее) значение показателя эффективности

$$F(\bar{x}) = \sum_{j=1}^{n} c_j x_j \to \max(\min)$$

при выполнении ограничений

$$\left.\begin{array}{l} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n \le b_1 \\ ..... \\ a_{p1}x_1 + a_{p2}x_2 + ... + a_{pn}x_n \le b_p \end{array}\right\}$$

$$\left.\begin{array}{l} a_{p+11}x_1 + a_{p+12}x_2 + ... + a_{p+1n}x_n \ge b_{p+1} \\ ..... \\ a_{q1}x_1 + a_{q2}x_2 + ... + a_{qn}x_n \ge b_q \end{array}\right\}$$

$$\left.\begin{array}{l} a_{q+11}x_1 + a_{q+12}x_2 + ... + a_{q+1n}x_n = b_{q+1} \\ ..... \\ a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n \le b_m \end{array}\right\}$$

и на некоторые переменные накладываются условия неотрицательности

$$x_1, x_2, ...x_r \ge 0, \quad r \le n \qquad (11)$$

Функция (7) называется целевой функцией или критерием оптимальности, или линейной формой.

Вектор управляемых параметров $\bar{x}$ называется решением. Решение называется допустимым, если оно удовлетворяет ограничениям (8–11). Допустимое решение называется планом.

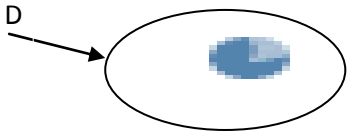
Решение $\bar{x}*$ называется оптимальным, если на нем достигается наибольшее значение критерия оптимальности $F(\bar{x})$:

$\bar{x}*$ – оптимальное решение, если $F(\bar{x}*) = \max F(\bar{x})$

$\bar{x}*$ – оптимальное решение, если для любого $\bar{x} \in D$, $F(\bar{x}) \le F(\bar{x}*)$ (13)

$\bar{x}* = Arg \max F(\bar{x})$

$\bar{x} \in D$



Задача линейного программирования называется разрешимой, если она имеет хотя бы одно оптимальное решение. У неразрешимой задачи или пуста область допустимых решений, или целевая функция не ограничена.

Различные формы задач линейного программирования

В зависимости от вида ограничений различают следующие формы задач:

- Каноническая
- Симметричная
- Общая.

Задача в канонической форме – задача ЛП, в которой все ограничения (8) – (10) есть равенства (p = q = 0) и все переменные неотрицательные (r = n).

$$F(\bar{x}) = \sum_{j=1}^{n} c_j x_j \to \max(\min)$$

$$\begin{cases} \sum_{j=1}^{n} a_{ij} x_j = b_i, \quad i = \overline{1,m} \\ x_j \ge 0, \quad j = \overline{1,n} \end{cases}$$

Общий метод решения задачи ЛП разработан именно для задачи в каноническом виде.

Матричный вид задачи в канонической форме:

$F(\bar{x}) = (\overline{c^t} * \bar{x}) \to \max$

$A\bar{x} = \bar{b}$

$\bar{x} \ge \bar{0}$

$\bar{c} = \begin{pmatrix} c_1 \\ c_2 \\ \dots \\ c_n \end{pmatrix}$ – вектор коэффициентов критерия

$\bar{x} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}$

$A = (a_{ij})_{m*n}$ – матрица условий (технологических коэффициентов)

$A = (\overline{A_1}, \overline{A_2}.....\overline{A_n})$

$\overline{A}_j = \begin{pmatrix} a_{1j} \\ a_{2j} \\ ... \\ a_{mj} \end{pmatrix}$ – вектор условий

$\overline{b} = \begin{pmatrix} b_1 \\ b_2 \\ ... \\ b_m \end{pmatrix}$ – вектор ограничений

$F(\overline{x}) = (\overline{c^t} * \overline{x}) \rightarrow \max$

$\overline{A}_1 x_1 + \overline{A}_2 x_2 + \ldots + \overline{A}_n x_n = \overline{b}$

$x_j \geq 0, \ j = \overline{1,n}$

Для неё разработан метод решения, который называется симплексным.

Задача в симметричной форме – задача ЛП, в которой все ограничения (8) - (10) есть неравенства (p = q = m) и все переменные неотрицательные (r = n).

$$F(\overline{x}) = \sum_{j=1}^{n} c_j x_j \rightarrow \max$$

$$\begin{cases} \sum_{j=1}^{n} a_{ij} x_j \leq b_i, \ \ i = \overline{1,m} \\ x_j \geq 0, \ \ j = \overline{1,n} \end{cases}$$

Матричный вид задачи в симметричной форме:

$$F(\overline{x}) = (\overline{c^t} * \overline{x}) \rightarrow \max$$

$$\begin{cases} A\overline{x} \leq \overline{b} \\ \overline{x} \geq \overline{0} \end{cases}$$

Симметричная форма допускает графическое решение (иллюстрацию).

Задача в общей (смешанной) форме – задача ЛП, в которой присутствуют все виды ограничений и не все переменные неотрицательные.

Графическое решение задач

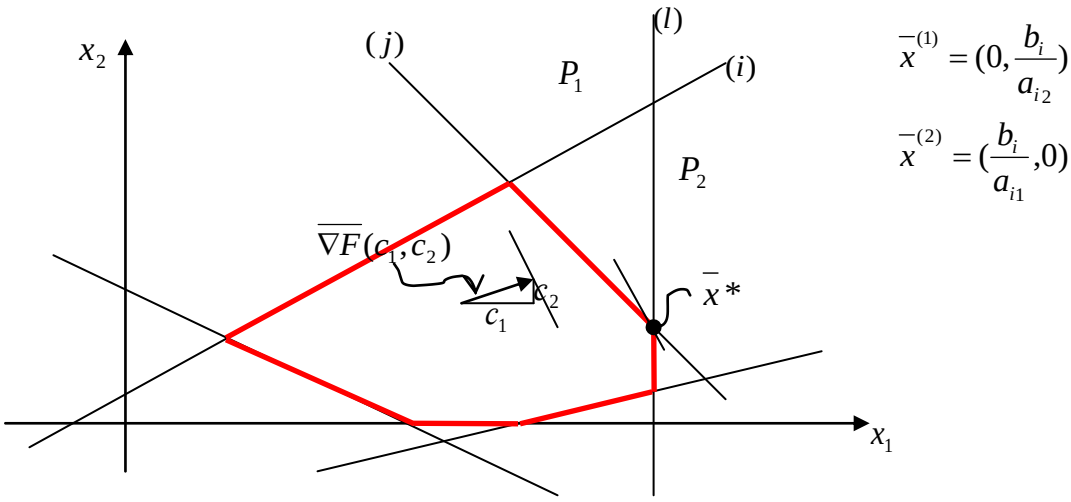
Задача ЛП в симметричной форме (17) – (19) может быть решена графически, если пространство управляемых параметров содержит две или три переменные.

Пусть $n = 2$, тогда задача ЛП в симметричной форме выглядит так:

$$F = c_1 x_1 + c_2 x_2 \rightarrow \max$$

$$\begin{cases} a_{i1} x_1 + a_{i2} x_2 \leq b_i, \ \ i = \overline{1,m} \\ x_1 \geq 0 \\ x_2 \geq 0 \end{cases}$$

Изобразим область допустимых решений в пространстве управляемых параметров. Для этого рассмотрим множество точек, удовлетворяющих неравенству (35). Это будет полуплоскость, ограниченная прямой.



Построим прямую $a_{i1}x_1 + a_{i2}x_2 = b_i$.

Эта прямая делит все пространство на две полуплоскости. Для определения допустимой полуплоскости выбираем произвольную точку пространства, не лежащую на прямой (37) (лучше всего точку (0,0)), и подставляем её координаты в ограничение (35). Если ограничение выполняется, то полуплоскость, содержащая эту точку, допустимая.

Пересечение плоскостей образует область допустимых решений D.

Поведение функции можно изобразить в пространстве управляемых параметров. Для этого достаточно построить семейство линий уровня функции. Линия уровня – это множество точек, в которых функция принимает постоянные значения.

$$F = c_1x_1 + c_2x_2 = const$$

Для задачи линейного программирования это прямые с разными константами в правой части уравнения (38).

Для построения линий уровня используем понятие градиента функции.

Градиент – это вектор частных производных функции.

$$\overline{grad}\ f(\bar{x}) = \overline{\nabla F} = (\frac{\partial f}{\partial x_1}; \frac{\partial f}{\partial x_2}; .... \frac{\partial f}{\partial x_n})$$

Его можно вычислить в любой конкретной точке пространства переменных.

Свойства градиента функции

- Градиент функции, вычисленный в точке, перпендикулярен линии уровня функции, проходящей через эту точку.
- Направление градиента показывает направление максимального возрастания функции.

Математические свойства задачи линейного программирования

Свойства области допустимых решений

Пусть дана задача в канонической форме:

$$F = \sum_{j=1}^{n} c_j x_j \to \max \qquad (1)$$

$$D\begin{cases} \sum_{j=1}^{n} a_{ij} x_j = b_i \quad i = \overline{1,m} \\ x_j \geq 0, \quad j = \overline{1,n} \end{cases} \qquad (2)$$

Пусть все уравнения линейно-независимые.

$rangA = m$

$\bar{c} \neq \bar{0}$

$D \neq \varnothing$

И пусть есть несколько n - мерных векторов $\bar{x}^{(1)}, \bar{x}^{(2)}, ...\bar{x}^{(r)}$.

Выпуклая оболочка n - мерных векторов – множество точек вида:

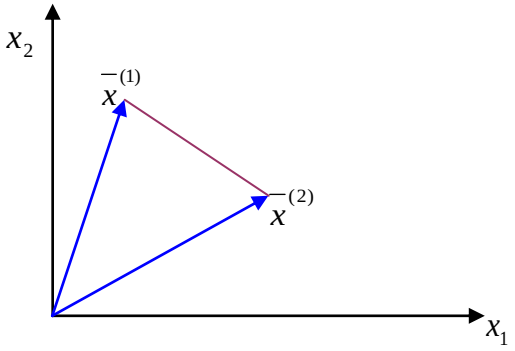
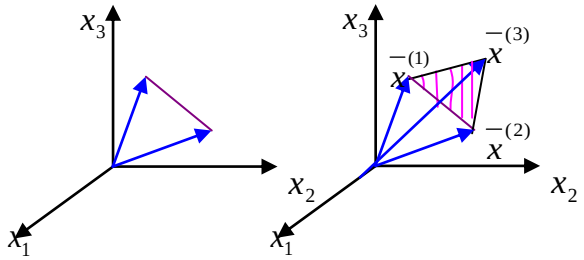
$$\bar{x} = \alpha_1 \bar{x}^{(1)} + \alpha_2 \bar{x}^{(2)} + ...\alpha_r \bar{x}^{(r)} \qquad (4)$$

$$\alpha_i \geq 0,\ i = \overline{1,r},\ \sum_{i=1}^{r} \alpha_i = 1$$

Выпуклая линейная комбинация двух векторов называется отрезком.

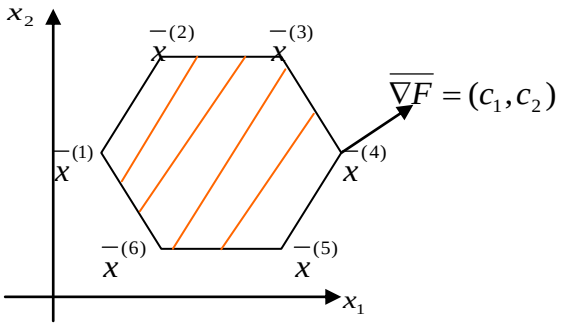
$\bar{x}^{(1)}, \bar{x}^{(2)}$

$$\bar{x} = \alpha \bar{x}^{(1)} + (1-\alpha)\bar{x}^{(2)},\ 0 \leq \alpha \leq 1 \qquad (6)$$

Двумерное пространство	Трехмерное пространство
	
$\bar{x} = \frac{1}{2}\bar{x}^{(1)} + \frac{1}{2}\bar{x}^{(2)}$	$\bar{x}^{(1)} = (3,5,10)$ $\bar{x}^{(2)} = (2,8,4)$ $\bar{x} = \alpha\begin{pmatrix}3\\5\\10\end{pmatrix} + (1-\alpha)\begin{pmatrix}2\\8\\4\end{pmatrix} = \begin{pmatrix}3\alpha\\5\alpha\\10\alpha\end{pmatrix} + \begin{pmatrix}2-2\alpha\\8-8\alpha\\4-4\alpha\end{pmatrix} = \begin{pmatrix}2+\alpha\\8-3\alpha\\4+6\alpha\end{pmatrix}$ $0 \leq \alpha \leq 1$

Область $D \subset R^n$ называется выпуклой, если вместе с любыми двумя своими точками она содержит отрезок, соединяющий их.

$$\begin{matrix} \forall \bar{x}^{(1)} \in D \\ \forall \bar{x}^{(2)} \in D \end{matrix} \quad \Rightarrow \quad \begin{matrix} \bar{x} = \alpha \bar{x}^{(1)} + (1-\alpha)\bar{x}^{(2)} \in D \\ 0 \leq \alpha \leq 1 \end{matrix}$$



Теорема 1: Область допустимых решений D задачи ЛП выпуклая.

Доказательство:

Пусть $\bar{x}^{(1)} \in D, \bar{x}^{(2)} \in D$, т.е. для них выполняются (2) и (3).

$$\bar{x} = \underset{\geq 0}{\alpha}\bar{x}^{(1)} + \underset{\geq 0}{(1-\alpha)}\bar{x}^{(2)} \qquad 0 \leq \alpha \leq 1$$

$$\Downarrow$$

$$\bar{x} \geq 0$$

таким образом, условие (3) выполняется.

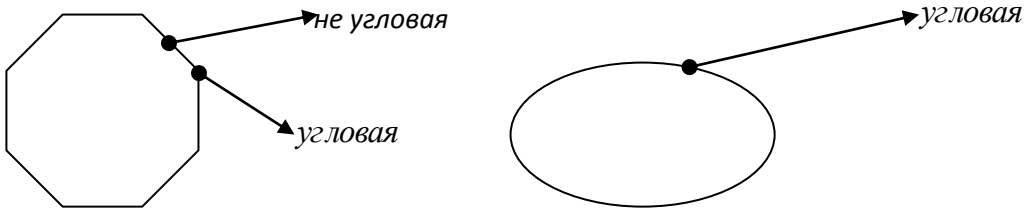
$$A\bar{x} = A(\alpha\bar{x}^{(1)} + (1-\alpha)\bar{x}^{(2)}) = \alpha A\bar{x}^{(1)} + (1-\alpha)A\bar{x}^{(2)} = \alpha\bar{b} + (1-\alpha)\bar{b} = \bar{b}$$

таким образом и условие (2) выполняется, произвольная точка отрезка принадлежит области D, то есть эта область выпукла.

Теорема доказана.

Точка $\bar{x} \in D \subset R^n$ называется угловой (крайней), если не существует двух других точек области, на отрезке между которыми лежит эта точка.

не $\exists\ \bar{x}' \in D$, $\bar{x}'' \in D$, $\bar{x}' \neq \bar{x}$, $\bar{x}'' \neq \bar{x}$, таких, что $\bar{x} = \alpha\bar{x}' + (1-\alpha)\bar{x}''$, $0 \leq \alpha \leq 1$ (7)



Лемма 1: Область допустимых решений D задачи ЛП является выпуклой линейной комбинацией своих угловых точек.

Таким образом, если найти все угловые точки, то любая точка внутри области записывается через уравнение (4).

Теорема 2: Оптимальное решение задачи ЛП достигается в одной из угловых точек области допустимых решений D.

Покажем, что оптимальное решение не может быть внутри области.

Пусть $\bar{x}^* = Arg\max F(\bar{x})$ – внутренняя точка области. Тогда функция дифференцируема в этой точке:

$$c_j = \frac{\partial f}{\partial x_j} = 0$$

Так как в этой точке достигается максимум, то и производная здесь обратится в ноль.

$\bar{c} = \bar{0}$

Но это противоречит условию, которое мы накладывали ранее, а именно $\bar{c} \neq \bar{0}$. Значит точка, в которой достигается оптимальное решение, лежит на границе области. Можно показать [8], что хотя бы одно оптимальное решение достигается в угловой точке.

Симплекс-метод решения задач ЛП

Этапы симплекс-метода

1. Проверка признака оптимальности ($\max \to \Delta_j \geq 0; \quad \min \to \Delta_j \leq 0$)

2. Если есть $\Delta_s < 0$, то решение не оптимальное. Тогда выбираем столбец с минимальной оценкой. Его назовем разрешающим.

3. Разрешающая строка выбирается по минимальному отношению свободных членов к положительным коэффициентам разрешающего столбца. Базисная переменная, выражающаяся из этой строки, выходит из списка базисных переменных. Т.е. x_k выходит, а x_s входит.

$$\frac{b_k^{|}}{a_{ks}^{|}} = \min_{i=1,m} \frac{b_i^{|}}{a_{is}^{|}}$$

$$a_{is}^{|} > 0$$

4. Текущая симплекс-таблица преобразуется по следующему правилу:

- разрешающая строка делится на разрешающий элемент:

$$a_{kj}^{н} = \frac{a_{kj}^{|}}{a_{ks}^{|}}$$

$$b_k^{н} = \frac{b_k^{|}}{a_{ks}^{|}}$$

- разрешающий столбец заменяется единичным.
- все остальные элементы симплекс-таблицы могут быть пересчитаны по правилу четырехугольника:

Мысленно строится четырехугольник на диагонали, соединяющей искомый элемент с разрешающим. Тогда новое значение элемента равно прежнему значению минус произведение элементов на противоположной диагонали, деленное на разрешающий элемент.

Или новое значение элемента равно произведению элементов на главной диагонали минус произведение элементов на противоположной диагонали, и все это деленное на разрешающий элемент.

$$a_{ij}^{н} = a_{ij}^{|} - \frac{a_{kj}^{|} a_{is}^{|}}{a_{ks}^{|}}$$

$$a_{ij}^{н} = \frac{a_{ij}^{|} a_{ks}^{|} - a_{kj}^{|} a_{is}^{|}}{a_{ks}^{|}}$$

2. Математическая формулировка двойственной задачи к произвольной задаче линейного программирования

Пусть исходная задача имеет вид:

$$F = \sum_{j=1}^{n} c_j x_j \to \max$$

$$\begin{cases} \sum_{j=1}^{n} a_{ij} x_j \le b_i, & i = \overline{1, p} \\ \sum_{j=1}^{n} a_{ij} x_j = b_i, & i = \overline{p+1, n} \\ x_j \ge 0, & j = \overline{1, r} \end{cases}$$

Тогда двойственной к задаче (7-10) называется задача вида:

$$T(\bar{y}) = \sum_{i=1}^{m} b_i y_i \to \min$$

$$\begin{cases} x_i \ge 0: & \sum_{i=1}^{m} a_{ij} y_i \le c_i, & i = \overline{1, r} \\ x_i \forall: & \sum_{i=1}^{m} a_{ij} y_i = c_i, & i = \overline{r+1, n} \\ y_i \ge 0, & i = \overline{1, p}, & y_i \forall, \quad i = \overline{p+1, m} \end{cases}$$

Правила построения двойственной задачи

Для применения правил, необходимо в задаче на максимум записать все ограничения – неравенства со знаком $\le$. В задаче же на минимум – со знаком $\ge$.

1. Количество переменных одной задачи совпадает с количеством ограничений другой задачи. Т.е. каждому ограничению одной задачи соответствует переменная другой. Ограничению-неравенству соответствует неотрицательная переменная, а ограничению-равенству – переменная произвольного знака.
2. Правые части ограничений одной задачи являются коэффициентами критерия другой.
3. Матрицы условий этих задач взаимно транспонированы, т.е. столбец матрицы условий одной задачи становится строкой другой.
4. Критерий одной задачи максимизируется, а другой минимизируется. Причем в задаче на максимум все ограничения – неравенства типа $\le$, а в задаче на минимум – типа $\ge$.

3. Нелинейное программирование

Задачи на условный экстремум. Метод множителей Лагранжа.

Задача на условный экстремум ставится как задача определения управляемых параметров $\bar{x} = (x_1, x_2, ...x_n)$, на которых достигается экстремум (максимум или минимум) при ограничениях, заданных уравнениями.

$$extr \ f(\bar{x}) \qquad (8)$$

$$g_i(x) = 0, \quad i = \overline{1, m}$$

$$m < n$$

Задачу на условный экстремум можно свести к задаче на безусловный экстремум для специальным образом построенной функции.

Каждому ограничению поставим в соответствие переменную λ_i.

Построим функцию Лагранжа:

$$L(\bar{x},\bar{\lambda}) = f(\bar{x}) + \sum_{i=1}^{m} \lambda_i g_i(x)$$

Если ограничения выполняются, то функция Лагранжа превращается в исходную функцию.

Точки локальных экстремумов задачи будут точками локальных экстремумов функции Лагранжа.

Задачи выпуклого программирования

Пусть исходная задача имеет вид:

$$\max f(x)$$

$$g_i(\bar{x}) \le b_i, \quad i = \overline{1,m}$$

$$x_j \ge 0, \quad j = \overline{1,n}$$

Задача (11-13) называется задачей выпуклого программирования, если выполняются следующие условия:

1. $f(\bar{x})$ – вогнутая функция, то есть для $\forall x^{(1)}, x^{(2)}$ $\quad f\left(\frac{x^{(1)} + x^{(2)}}{2}\right) \ge \frac{f(x^{(1)}) + f(x^{(2)})}{2}$
2. область допустимых решений D выпуклая
3. область регулярная, то есть существует по крайней мере одна внутренняя точка $g_i(x') < b_i$

Можно построить функцию Лагранжа:

$$L(\bar{x},\bar{\lambda}) = f(x) + \sum_{i=1}^{m} \lambda_i (b_i - g_i(\bar{x}))$$

Теорема 8: точка $\bar{x}^*$ является оптимальным решением задачи выпуклого программирования тогда и только тогда, когда существует вектор $\bar{\lambda}^*$ такой, что для функции Лагранжа выполняются условия:

1. $$\frac{\partial L(\bar{x}^*,\bar{\lambda}^*)}{\partial x_j} \le 0, \quad j = \overline{1,n}$$
   $$\frac{\partial L(\bar{x}^*,\bar{\lambda}^*)}{\partial \lambda_i} \ge 0, \quad i = \overline{1,m}$$
2. условия дополняющей нежесткости:

$$x_j^* \times \frac{\partial L(\bar{x}^*,\bar{\lambda}^*)}{\partial x_j} = 0, \quad j = \overline{1,n}$$

$$\lambda_i^* \times \frac{\partial L(\bar{x}^*,\bar{\lambda}^*)}{\partial \lambda_i} = 0, \quad i = \overline{1,m}$$

3. $x_j^* \geq 0, \quad j = \overline{1,n}$

$\lambda_i^* \geq 0, \quad i = \overline{1,m}$

Эти условия называются условиями Куна – Таккера.

В общем случае система уравнений и неравенств (14) слишком сложна для аналитического решения. Однако в задачах квадратичного программирования есть способы решения этой системы условий, сводящиеся к нахождению опорных решений систем линейных алгебраических уравнений.

Геометрический смысл условий Куна-Таккера:

Если все $x_j^x > 0$, $j = \overline{1,n}$, то условия дополняющей нежесткости запишутся в виде

$$\nabla f(\overrightarrow{x^x}) = \sum_{i=1}^{m} \lambda_i \nabla g_i(\overrightarrow{x^x})$$

$$g_i(\overrightarrow{x^x}) = b_i$$

то есть градиент функции $f(\vec{x})$ в оптимальной точке является линейной комбинацией с положительными коэффициентами градиентов и активным ограничениям. Иными словами, градиент критерия лежит в геометрическом конусе градиентов ограничений.

Вопросы по лекции

1. Назначение и область применения линейного программирования.
2. Общий вид задачи ЛП.
3. Применения линейного программирования в принятии решений.
4. Формулировка общей задачи математического программирования.
5. Седловая точка и методы ее поиска.
6. Формулировка теоремы Кунна-Таккера.
7. Алгоритм применения графического метода решения задачи нелинейного программирования.
8. Особенности градиентного метода решения задач математического программирования.
9. Особенности применения метода возможных направлений в моделировании.
10. Понятие и область применение метода Монте-Карло.

Практикум по модулю 2

Практическое занятие №5 Прогнозирование на основе адаптивных моделей

По предложенному варианту составить прогноз на три шага вперед, используя:

а) Модель прогнозной экстраполяции

б) Модель авторегрессии

в) Модель авторегрессии-проинтегрированного скользящего среднего

г) Метод экспоненциального сглаживания.

n - момент наблюдения, x - наблюдаемый параметр.

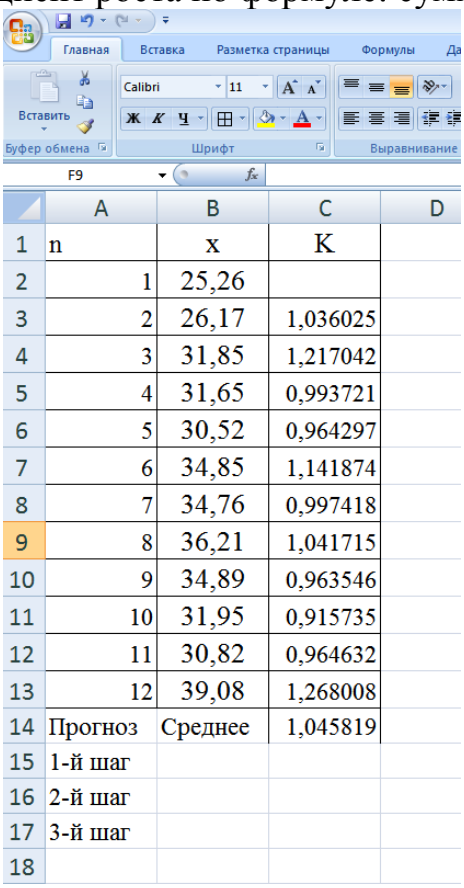
Пример решения задачи

Исходные данные

n	1	2	3	4	5	6	7	8	9	10	11	12
x	25,26	26,17	31,85	31,65	30,52	34,85	34,76	36,21	34,89	31,95	30,82	39,08

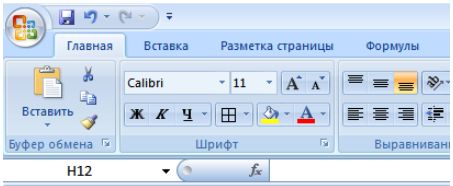
1. Модель прогнозной экстраполяции

Будем использовать экстраполяцию по среднемук коэффициенту роста. Рассчитаем средний коэффициент роста по формуле: сумма(x_i/x_{i-1})/n



	A	B	C	D
1	n	x	K	
2	1	25,26		
3	2	26,17	1,036025	
4	3	31,85	1,217042	
5	4	31,65	0,993721	
6	5	30,52	0,964297	
7	6	34,85	1,141874	
8	7	34,76	0,997418	
9	8	36,21	1,041715	
10	9	34,89	0,963546	
11	10	31,95	0,915735	
12	11	30,82	0,964632	
13	12	39,08	1,268008	
14	Прогноз	Среднее	1,045819	
15	1-й шаг			
16	2-й шаг			
17	3-й шаг			
18				

Для расчета прогноза будем использовать формулу: $x_{прог}=x_{баз}*K^t_{сред}$ - где t - шаг прогнозирования.



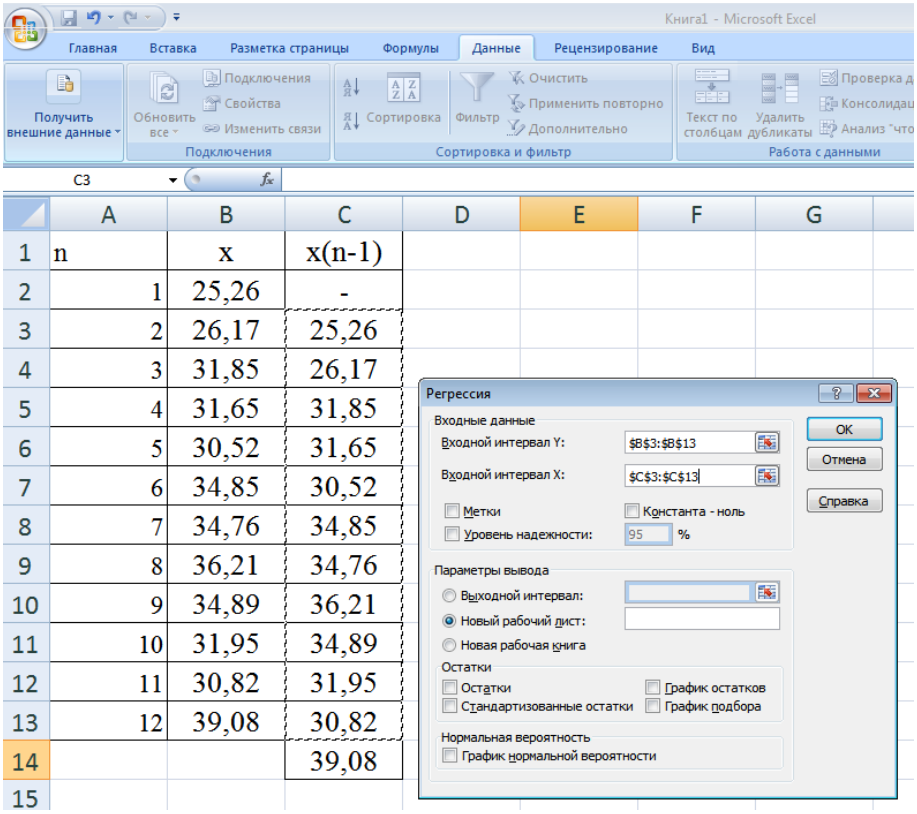
	A	B	C	D
1	n	x	K	
2	1	25,26		
3	2	26,17	1,036025	
4	3	31,85	1,217042	
5	4	31,65	0,993721	
6	5	30,52	0,964297	
7	6	34,85	1,141874	
8	7	34,76	0,997418	
9	8	36,21	1,041715	
10	9	34,89	0,963546	
11	10	31,95	0,915735	
12	11	30,82	0,964632	
13	12	39,08	1,268008	
14	Прогноз	Среднее	1,045819	
15	1-й шаг	33,8157		
16	2-й шаг	35,36511		
17	3-й шаг	36,98552		

2. Модель авторегрессии.

Подготовим исходные данные:

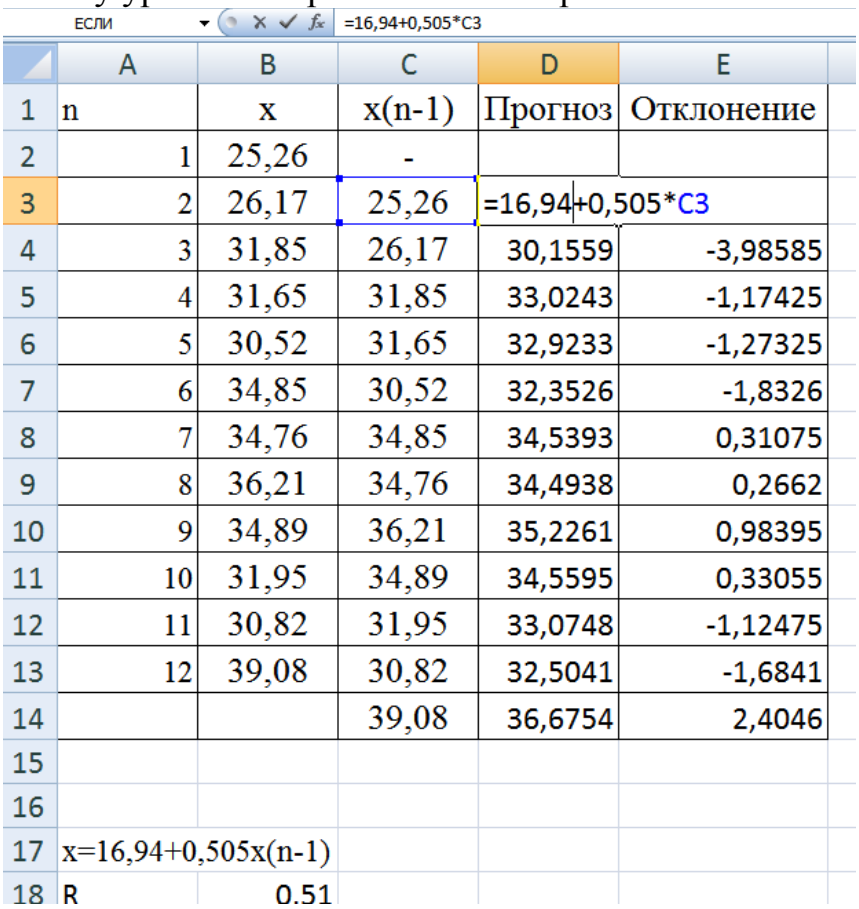
	A	B	C
1	n	x	x(n-1)
2	1	25,26	-
3	2	26,17	25,26
4	3	31,85	26,17
5	4	31,65	31,85
6	5	30,52	31,65
7	6	34,85	30,52
8	7	34,76	34,85
9	8	36,21	34,76
10	9	34,89	36,21
11	10	31,95	34,89
12	11	30,82	31,95
13	12	39,08	30,82
14			39,08
15			

Потроим уравнение регрессии:



	A	B	C
1	n	x	x(n-1)
2	1	25,26	-
3	2	26,17	25,26
4	3	31,85	26,17
5	4	31,65	31,85
6	5	30,52	31,65
7	6	34,85	30,52
8	7	34,76	34,85
9	8	36,21	34,76
10	9	34,89	36,21
11	10	31,95	34,89
12	11	30,82	31,95
13	12	39,08	30,82
14			39,08
15			

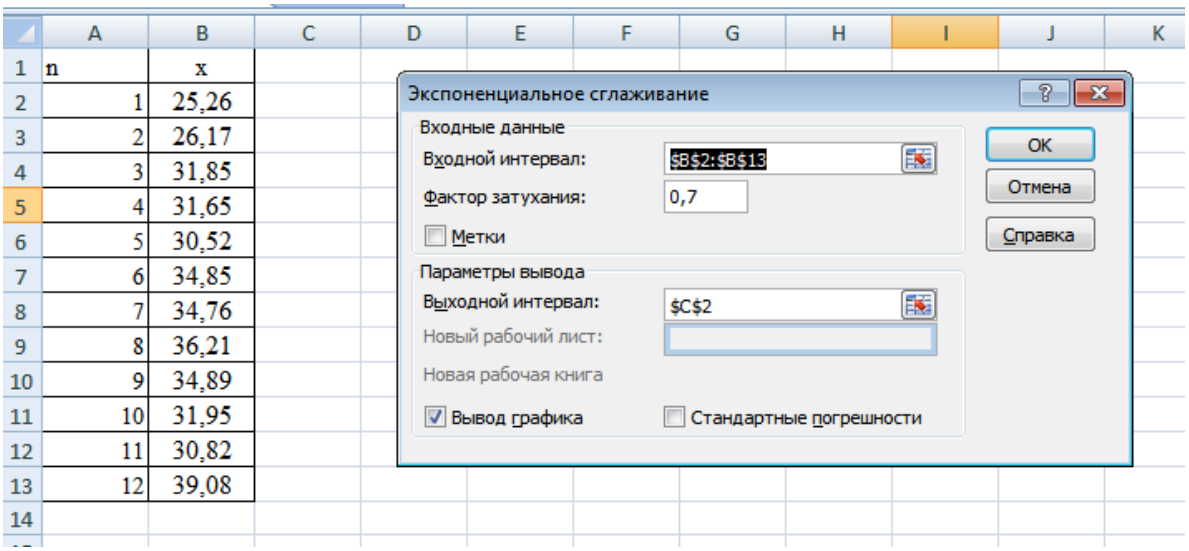
По построенному уравнению рассчитываем прогнозные значения:



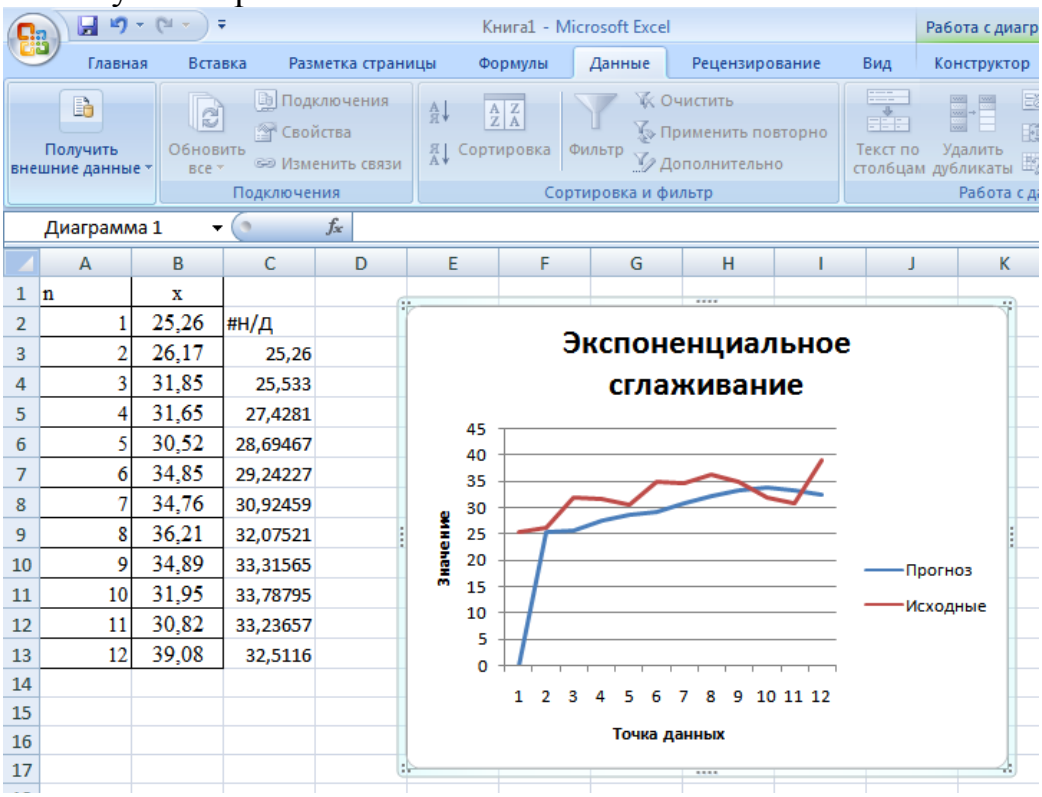
	A	B	C	D	E
1	n	x	x(n-1)	Прогноз	Отклонение
2	1	25,26	-		
3	2	26,17	25,26	=16,94+0,505*C3	
4	3	31,85	26,17	30,1559	-3,98585
5	4	31,65	31,85	33,0243	-1,17425
6	5	30,52	31,65	32,9233	-1,27325
7	6	34,85	30,52	32,3526	-1,8326
8	7	34,76	34,85	34,5393	0,31075
9	8	36,21	34,76	34,4938	0,2662
10	9	34,89	36,21	35,2261	0,98395
11	10	31,95	34,89	34,5595	0,33055
12	11	30,82	31,95	33,0748	-1,12475
13	12	39,08	30,82	32,5041	-1,6841
14			39,08	36,6754	2,4046
15					
16					
17	x=16,94+0,505x(n-1)				
18	R	0,51			

3. Модель экспоненциального сглаживания

Выберем в менб Данные надстройку Анализ данных и инструмент Экспоненциальное сглаживание, введя исходный ряд и установив фактор затухания 0,7:



В итоге получаем прогноз:



Варианты

Вариант 1

n	1	2	3	4	5	6	7	8	9	10	11	12
x	20,26	29,17	28,85	31,65	24,52	31,85	32,76	29,21	28,89	32,95	30,82	19,08

Вариант 2

n	1	2	3	4	5	6	7	8	9	10	11	12
x	13,00	15,00	20,00	24,00	27,50	31,30	35,10	38,90	42,70	46,50	50,30	54,10

Вариант 3

n	1	2	3	4	5	6	7	8	9	10	11	12
x	24,00	36,00	29,00	43,00	42,00	47,70	52,00	56,30	60,60	64,90	69,20	73,50

Вариант 4

n	1	2	3	4	5	6	7	8	9	10	11	12
x	10,00	11,00	13,00	15,00	13,00	14,00	15,47	16,27	17,07	17,87	18,67	19,47

Вариант 5

n	1	2	3	4	5	6	7	8	9	10	11	12
x	25,00	27,00	28,00	31,00	28,00	30,80	31,80	32,80	33,80	34,80	35,80	36,80

Вариант 6

n	1	2	3	4	5	6	7	8	9	10	11	12
x	30,00	35,00	43,00	41,00	39,00	30,80	37,87	38,27	38,67	39,07	39,47	39,87

Вариант 7

n	1	2	3	4	5	6	7	8	9	10	11	12
x	63,00	63,00	66,00	63,00	59,50	60,80	60,10	59,40	58,70	58,00	57,30	56,60

Вариант 8

n	1	2	3	4	5	6	7	8	9	10	11	12
x	11,00	15,00	21,00	19,00	18,00	22,20	24,00	25,80	27,60	29,40	31,20	33,00

Вариант 9

n	1	2	3	4	5	6	7	8	9	10	11	12
x	14,00	18,00	19,00	23,00	25,50	28,30	31,10	33,90	36,70	39,50	42,30	45,10

Вариант 10

n	1	2	3	4	5	6	7	8	9	10	11	12
x	19,00	20,00	30,00	25,00	27,00	30,50	32,60	34,70	36,80	38,90	41,00	43,10

Практическое занятие № 6 Построение прогноза по нейросети.

Используя ППП Statistica, SPSS или Deductor построить и обучить нейронную сеть на основе однослойного персептрона.

X1, X2 … X8 некоторые признаки, характеризующие исследуемый процесс. Выходной признак X1, остальные признаки входные.

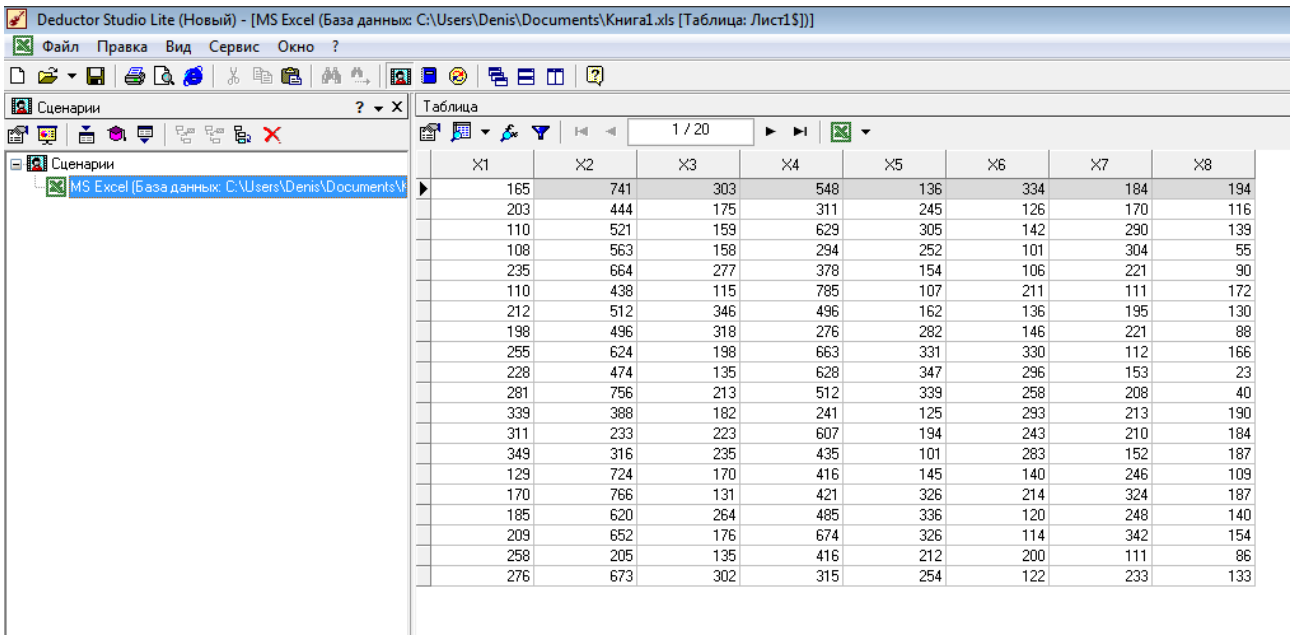
Пример решения задачи

n	X1	X2	X3	X4	X5	X6	X7	X8
1	165	741	303	548	136	334	184	194
2	203	444	175	311	245	126	170	116
3	110	521	159	629	305	142	290	139
4	108	563	158	294	252	101	304	55
5	235	664	277	378	154	106	221	90
6	110	438	115	785	107	211	111	172
7	212	512	346	496	162	136	195	130
8	198	496	318	276	282	146	221	88
9	255	624	198	663	331	330	112	166
10	228	474	135	628	347	296	153	23
11	281	756	213	512	339	258	208	40
12	339	388	182	241	125	293	213	190
13	311	233	223	607	194	243	210	184
14	349	316	235	435	101	283	152	187
15	129	724	170	416	145	140	246	109
16	170	766	131	421	326	214	324	187
17	185	620	264	485	336	120	248	140
18	209	652	176	674	326	114	342	154
19	258	205	135	416	212	200	111	86
20	276	673	302	315	254	122	233	133

1. Определим интервалы изменения значений признаков:

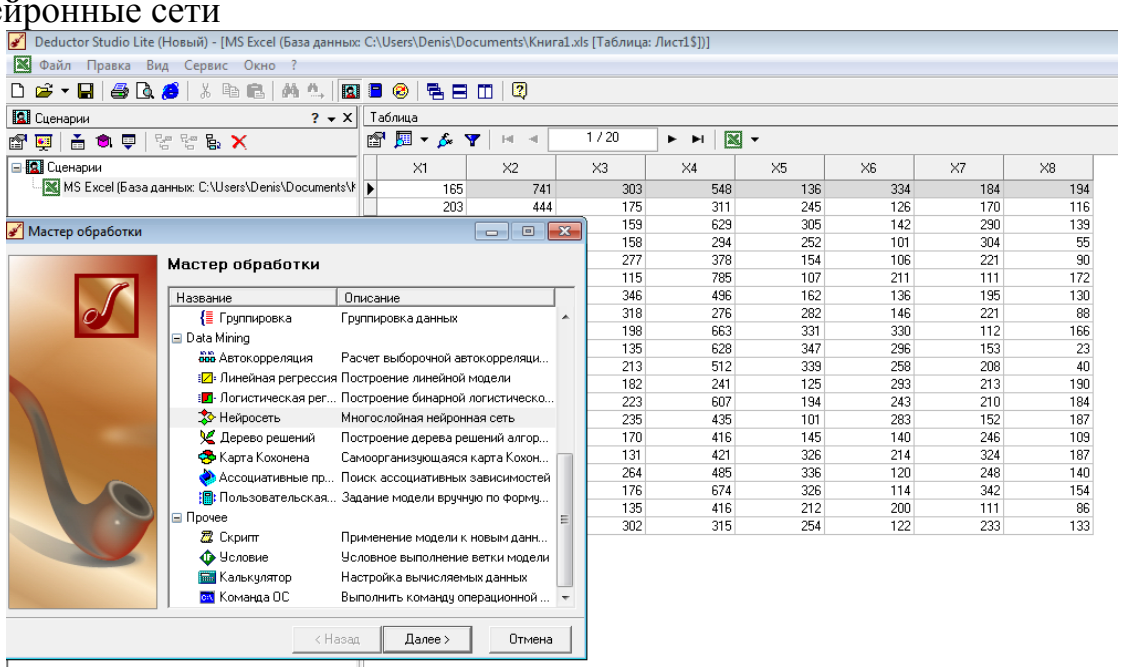
	X1	X2	X3	X4	X5	X6	X7	X8
Мин	108	205	115	241	101	101	111	23
Макс	349	766	346	785	347	334	342	194

Воспользуемся платформой Deductor и импортируем исходные данные

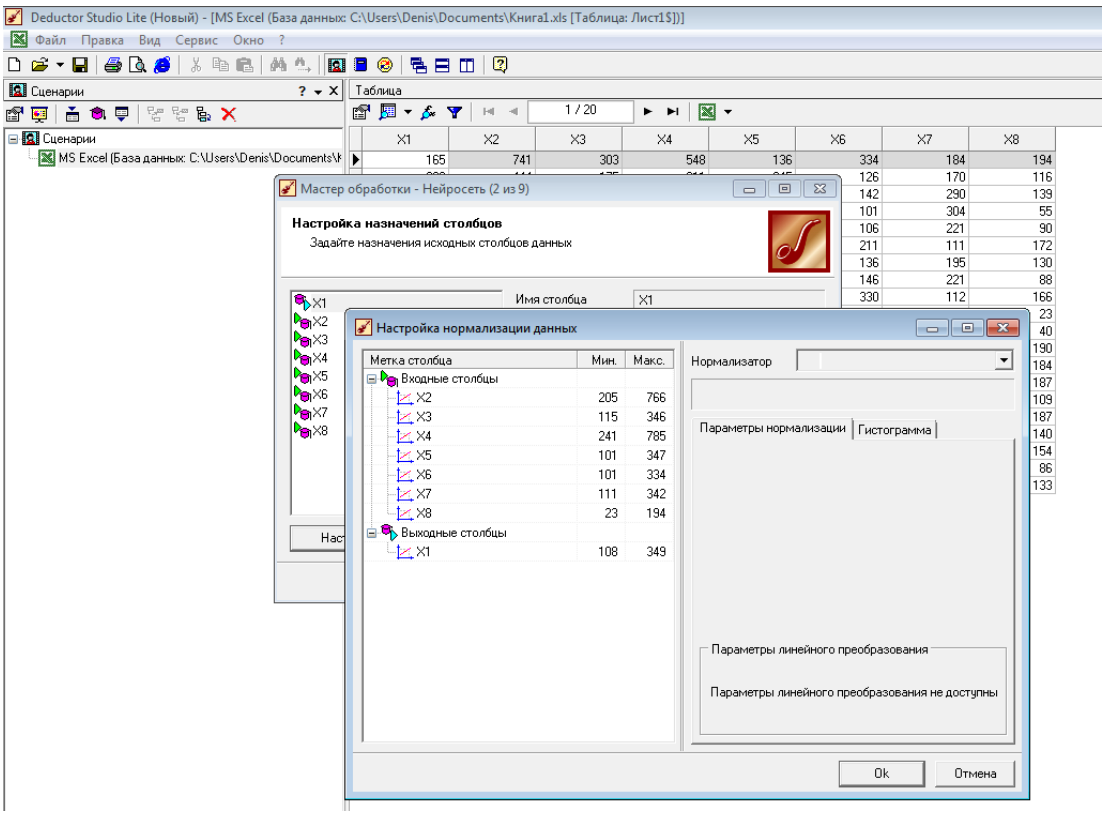


X1	X2	X3	X4	X5	X6	X7	X8
165	741	303	548	136	334	184	194
203	444	175	311	245	126	170	116
110	521	159	629	305	142	290	139
108	563	158	294	252	101	304	55
235	664	277	378	154	106	221	90
110	438	115	785	107	211	111	172
212	512	346	496	162	136	195	130
198	496	318	276	282	146	221	88
255	624	198	663	331	330	112	166
228	474	135	628	347	296	153	23
281	756	213	512	339	258	208	40
339	388	182	241	125	293	213	190
311	233	223	607	194	243	210	184
349	316	235	435	101	283	152	187
129	724	170	416	145	140	246	109
170	766	131	421	326	214	324	187
185	620	264	485	336	120	248	140
209	652	176	674	326	114	342	154
258	205	135	416	212	200	111	86
276	673	302	315	254	122	233	133

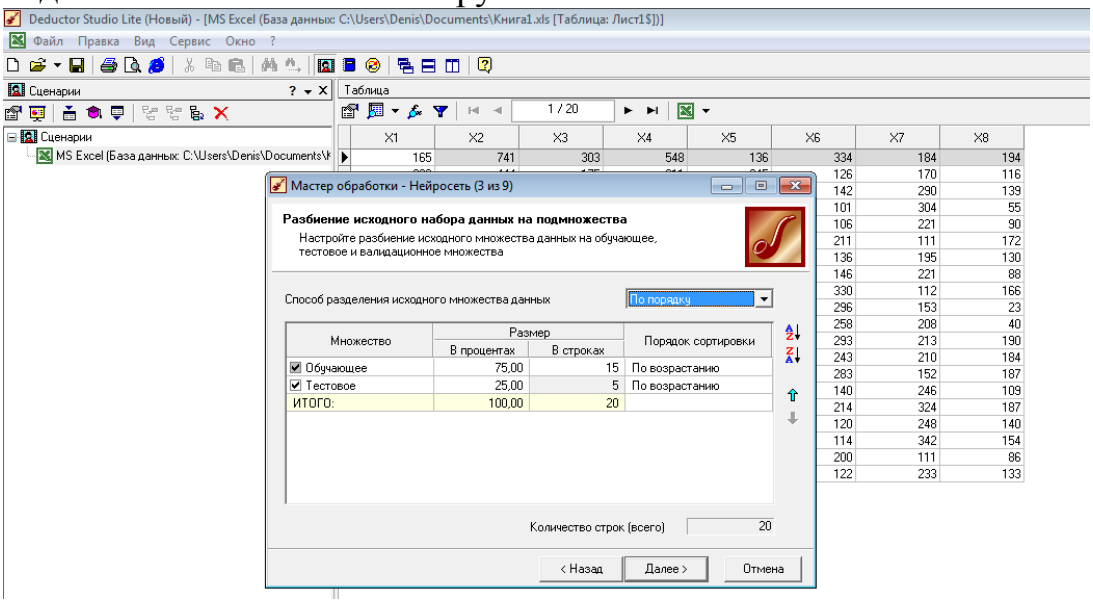
2. Используя мастер обработки, в разделе data mining выбираем нейронные сети



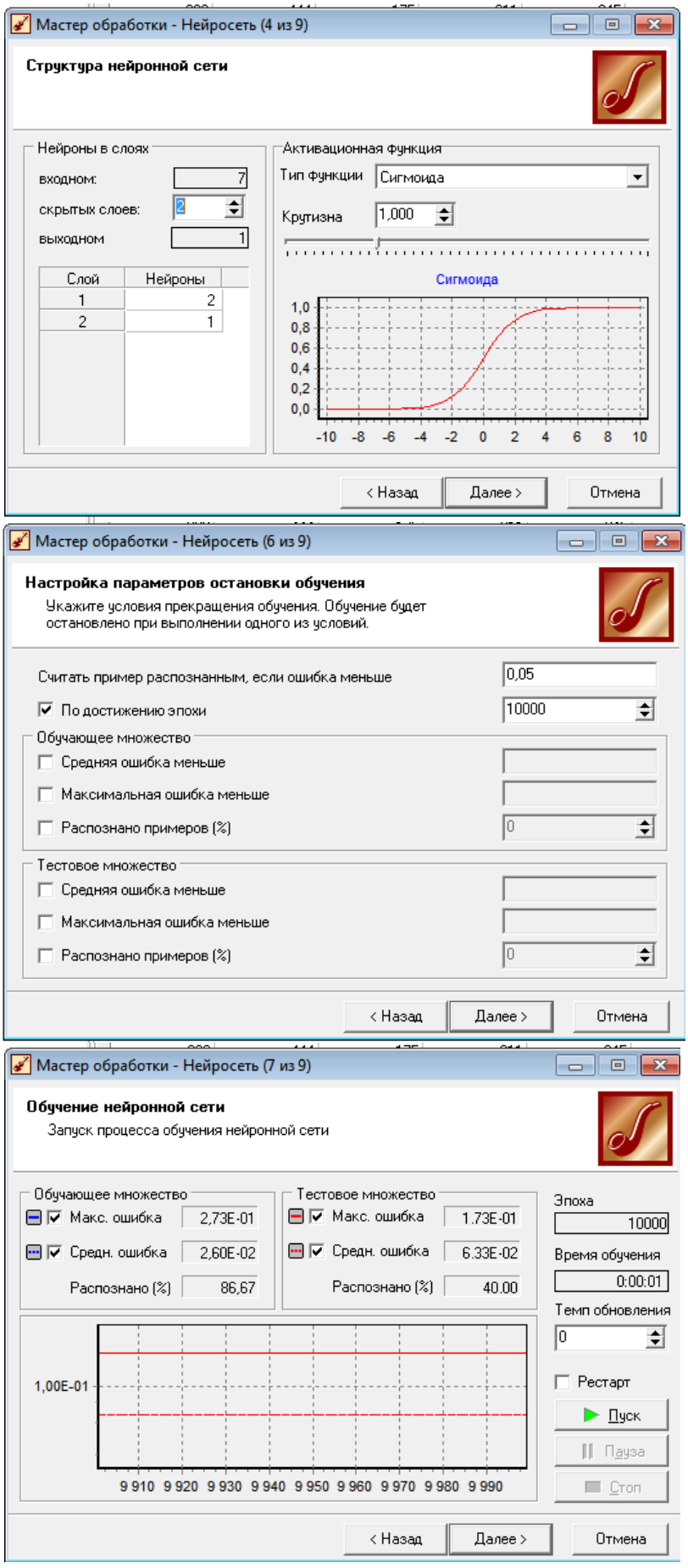
3. В мастере обработки показатель X1 делаем выходным, а остальные входные:



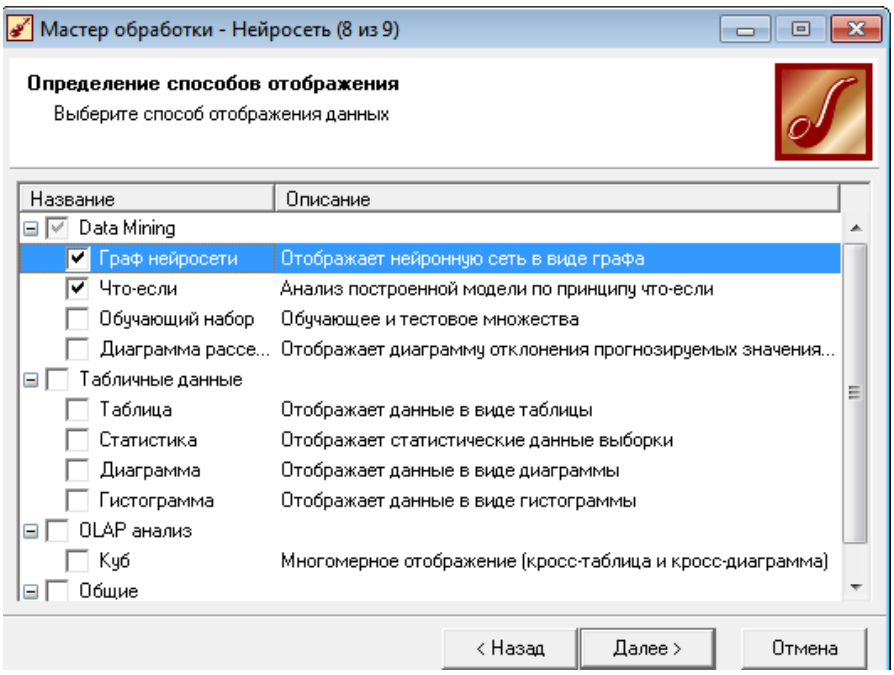
Первые 15 значений устанавливаем в качестве обучаемого множества, а последние 5 значений как тестируемое множество:



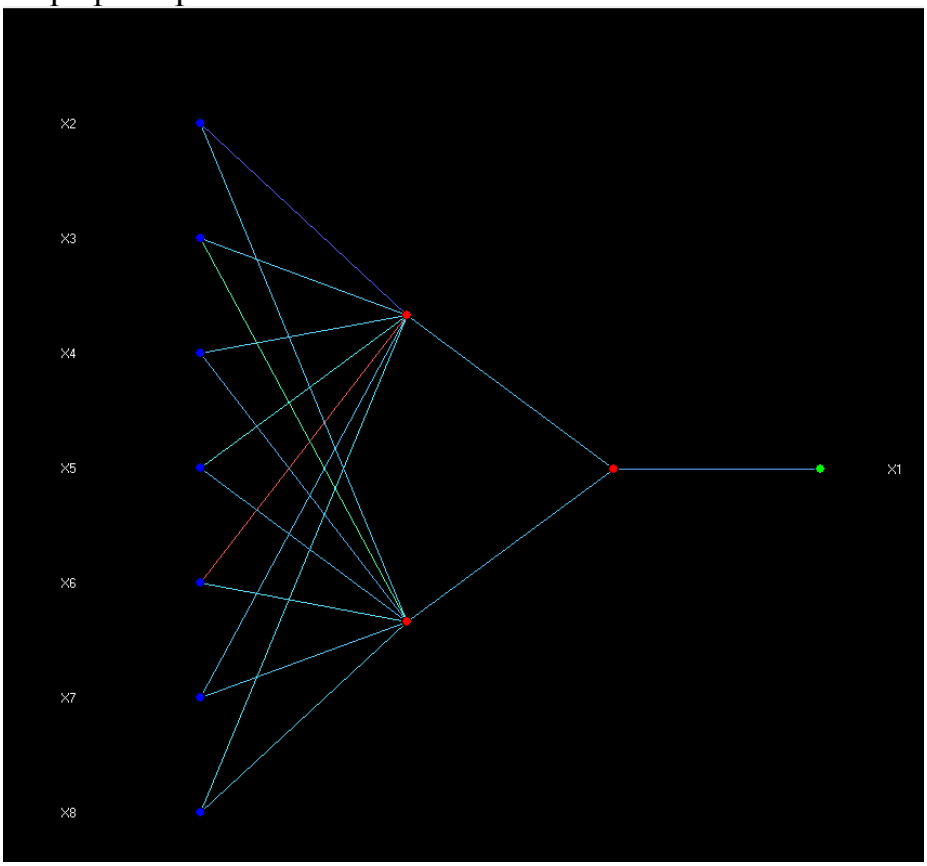
4. Далее выбираем сигмоидальную функцию и два скрытых слоя с 5% ошибкой запускаем обучение нейросети:



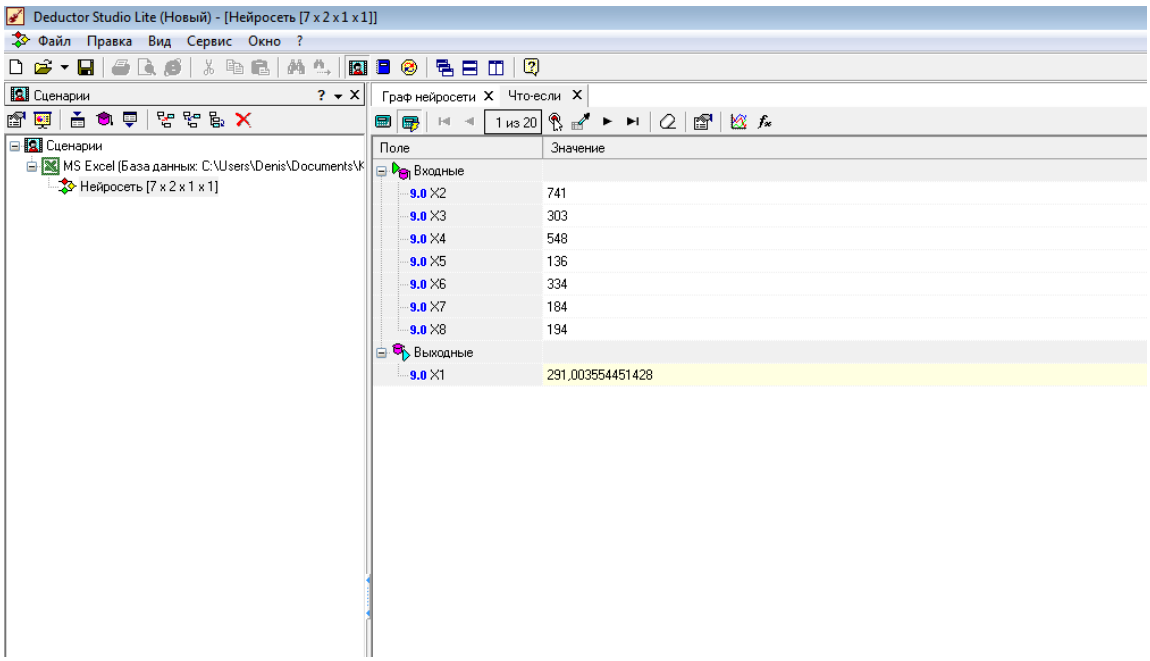
5. Для того, чтобы иметь возможность прогнозирования устанавливаем галочку на анализе «что-если»:



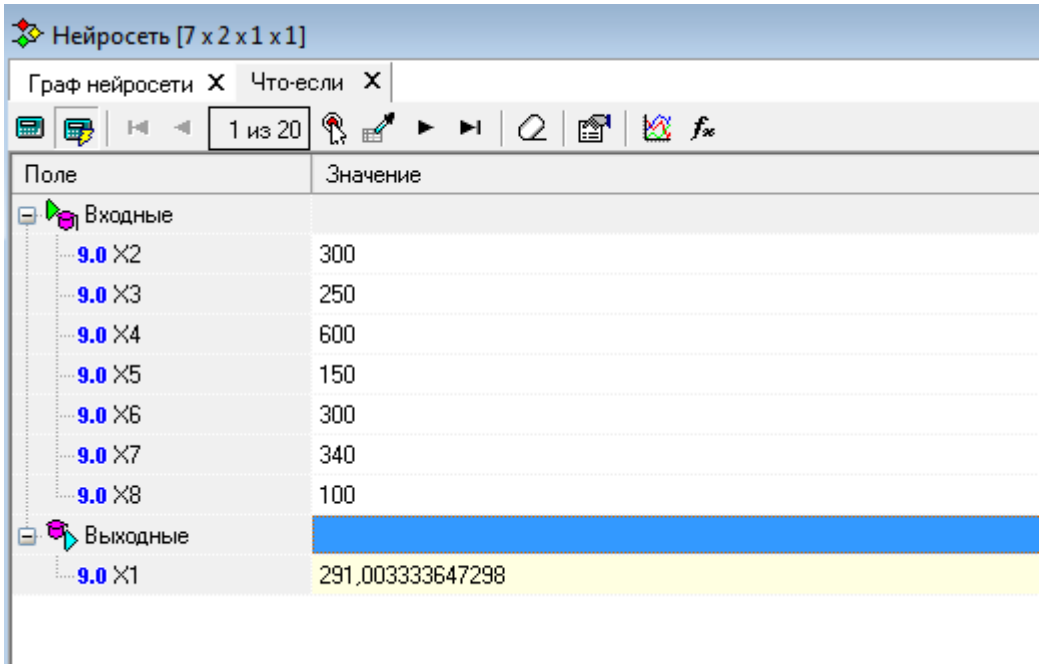
Получаем граф нейросети:



И анализ «что-если»:



6. Теперь устанавливая значения для признаков X2-X8 из интервалов их значений, рассчитаных на 1-м шаге, мы можем получить прогнозное значение признака X1. Например, если X2=300, X3=250, X4=600, X5=150, X6=300, X7=340, X8=100, то X1 будем равно 291 с 5% ошибкой:



Варианты

Вариант 1

n	X1	X2	X3	X4	X5	X6	X7	X8
1	165	119	303	342	136	334	184	128
2	203	222	175	182	245	126	170	167
3	110	132	159	242	305	142	290	286
4	108	202	158	179	252	101	304	214
5	235	104	277	325	154	106	221	211
6	110	311	115	304	107	211	111	133
7	212	244	346	307	162	136	195	161
8	198	315	318	276	282	146	221	188
9	255	132	198	234	331	330	112	103
10	228	113	135	284	347	296	153	243
11	281	249	213	257	339	258	208	242
12	339	275	182	264	125	293	213	293
13	311	256	223	166	194	243	210	310
14	349	315	235	183	101	283	152	248
15	129	280	170	174	145	140	246	336
16	170	227	131	316	326	214	324	208
17	185	191	264	193	336	120	248	190
18	209	267	176	284	326	114	342	251
19	258	136	135	119	212	200	111	102
20	276	230	302	330	254	122	233	210

Вариант 2

n	X1	X2	X3	X4	X5	X6	X7	X8
1	267	741	456	548	323	671	237	514
2	639	444	660	311	797	308	701	536
3	707	521	521	629	798	266	272	710
4	466	563	755	294	526	606	386	259
5	665	664	645	378	627	418	671	602
6	341	438	384	785	515	478	581	453
7	716	512	201	496	654	570	205	598
8	478	496	454	276	428	501	250	342
9	431	624	303	663	582	575	601	632
10	722	474	352	628	630	353	454	661
11	246	756	278	512	253	478	623	536
12	622	388	417	241	458	387	226	432
13	344	233	745	607	321	286	342	413
14	710	316	215	435	700	209	244	335
15	723	724	687	416	527	346	612	416
16	772	766	417	421	669	249	510	688
17	677	620	530	485	374	516	750	520

18	526	652	468	674	678	340	421	618
19	668	205	411	416	401	545	403	299
20	296	673	469	315	312	690	645	239

Вариант 3

n	X1	X2	X3	X4	X5	X6	X7	X8
1	120	122	52	189	166	91	105	194
2	55	56	53	109	108	23	111	116
3	95	23	108	151	37	51	140	139
4	151	90	66	77	196	111	172	55
5	33	27	148	97	59	43	140	90
6	75	117	137	34	99	50	172	172
7	84	185	135	140	62	96	50	130
8	189	93	157	133	63	56	30	88
9	197	71	28	52	172	58	172	166
10	126	181	110	63	104	133	123	23
11	143	152	160	97	118	70	111	40
12	70	147	100	62	31	184	53	190
13	78	105	184	65	156	138	54	184
14	62	190	190	129	190	121	105	187
15	22	183	172	102	188	68	78	109
16	20	154	62	137	44	164	194	187
17	67	75	100	106	146	113	165	140
18	124	182	126	126	179	47	63	154
19	36	92	51	197	193	193	50	86
20	117	108	97	53	165	33	30	133

Вариант 4

n	X1	X2	X3	X4	X5	X6	X7	X8
1	156	111	117	79	44	96	144	127
2	122	172	178	60	78	52	111	121
3	181	187	148	77	43	46	117	144
4	110	195	179	98	72	63	103	103
5	110	117	187	66	41	84	143	118
6	165	136	160	83	67	87	147	115
7	154	189	147	55	83	54	121	122
8	181	106	149	77	40	81	138	139
9	127	154	199	52	62	54	109	137
10	104	157	191	71	60	56	147	133
11	163	108	192	97	64	46	142	145
12	141	178	198	74	93	98	135	144

13	143	135	192	59	42	85	130	105
14	143	116	100	94	45	50	131	118
15	153	169	163	91	42	61	140	108
16	102	137	189	93	48	61	120	108
17	196	135	135	100	43	81	127	134
18	136	115	128	75	49	90	147	109
19	113	126	108	85	73	83	144	140
20	173	171	163	61	50	72	129	128

Вариант 5

n	X1	X2	X3	X4	X5	X6	X7	X8
1	117	122	62	55	418	271	252	346
2	140	133	69	42	448	267	480	335
3	144	122	45	70	471	212	247	327
4	142	137	35	61	471	357	279	335
5	111	103	64	43	474	380	478	328
6	128	106	33	36	467	244	396	345
7	128	120	37	39	412	474	387	322
8	115	126	53	59	466	499	373	320
9	101	130	64	37	408	266	293	339
10	111	148	51	54	420	213	308	393
11	122	147	58	57	471	325	316	312
12	147	104	64	34	401	420	231	325
13	105	126	45	65	466	421	342	394
14	137	123	49	58	439	454	241	339
15	117	129	61	41	472	384	210	352
16	129	144	62	32	489	410	297	307
17	119	112	30	33	422	365	442	394
18	111	150	40	42	417	298	493	340
19	148	134	61	32	456	428	369	359
20	131	112	43	38	400	496	491	337

Вариант 6

n	X1	X2	X3	X4	X5	X6	X7	X8
1	331	345	376	44	26	136	139	116
2	340	387	360	57	34	114	148	133
3	363	345	396	40	29	143	126	100
4	307	371	360	36	51	140	111	118
5	366	349	395	59	20	139	112	107
6	306	376	373	46	26	131	122	128
7	371	333	342	58	27	144	113	111
8	384	352	321	39	52	142	135	113
9	371	375	370	54	21	134	131	144

10	390	391	348	51	36	123	130	110
11	333	336	309	20	58	125	123	102
12	334	320	351	22	26	120	124	139
13	317	307	309	39	34	115	141	134
14	347	383	322	46	37	120	113	115
15	323	345	330	26	23	149	142	119
16	378	300	360	56	31	131	149	148
17	368	343	397	29	48	111	112	116
18	321	376	322	26	51	131	147	101
19	319	371	380	43	56	147	117	150
20	330	393	387	60	25	103	135	108

Вариант 7

n	X1	X2	X3	X4	X5	X6	X7	X8
1	148	112	131	673	600	579	51	52
2	118	134	117	493	561	597	53	44
3	116	126	138	516	675	700	54	56
4	110	131	113	628	562	667	40	53
5	139	133	108	705	565	711	45	56
6	135	108	126	404	768	595	49	44
7	129	111	141	360	568	463	48	59
8	105	119	123	370	669	578	57	44
9	105	118	112	604	506	383	48	53
10	115	104	105	673	788	500	57	50
11	148	111	124	389	311	531	53	52
12	126	126	100	397	523	319	42	48
13	143	129	125	619	795	381	57	51
14	102	125	132	505	579	493	47	43
15	148	121	133	663	693	401	53	57
16	132	110	144	391	318	367	59	55
17	127	121	145	408	439	484	60	45
18	139	139	120	393	412	735	40	47
19	114	113	133	515	508	545	47	60
20	104	120	146	592	594	545	60	52

Вариант 8

n	X1	X2	X3	X4	X5	X6	X7	X8
1	44	42	166	181	330	392	39	55
2	44	42	243	106	430	201	51	43
3	47	46	203	255	287	456	58	51
4	45	54	113	171	332	225	29	29
5	58	57	117	225	397	288	16	24
6	56	50	238	107	262	406	27	20

7	44	56	150	151	273	444	26	60
8	49	49	237	235	203	326	19	52
9	45	53	201	200	395	219	21	42
10	50	60	249	150	243	448	37	15
11	43	54	149	227	224	263	39	30
12	46	60	105	232	288	271	33	41
13	49	51	194	139	296	285	16	53
14	41	59	198	227	322	325	55	48
15	56	59	106	104	327	253	29	53
16	52	56	125	113	333	253	20	49
17	54	60	142	231	372	303	48	57
18	55	41	247	220	317	456	55	55
19	50	52	107	158	269	261	20	52
20	59	49	160	202	301	426	24	37

Вариант 9

n	X1	X2	X3	X4	X5	X6	X7	X8
1	41	40	166	181	330	1872	1772	1196
2	18	25	243	106	430	1642	1377	1777
3	58	55	203	255	287	1798	1465	1586
4	55	30	113	171	332	1419	1141	1388
5	38	50	117	225	397	1677	1404	1904
6	17	49	238	107	262	1494	1562	1015
7	50	52	150	151	273	1661	1269	1631
8	53	41	237	235	203	1023	1791	1021
9	50	56	201	200	395	1140	1637	1233
10	18	42	249	150	243	1724	1281	1419
11	35	47	149	227	224	1381	1471	1471
12	45	39	105	232	288	1552	1889	1168
13	22	49	194	139	296	1875	1386	1681
14	43	28	198	227	322	1643	1711	1644
15	26	37	106	104	327	1932	1820	1364
16	54	46	125	113	333	1223	1315	1105
17	56	17	142	231	372	1832	1967	1285
18	22	43	247	220	317	1663	1542	1344
19	41	19	107	158	269	1686	1713	1068
20	45	43	160	202	301	1416	1465	1524

Вариант 10

n	X1	X2	X3	X4	X5	X6	X7	X8
1	1734	1041	1725	148	172	152	19	14
2	1347	1497	1375	137	199	124	38	17
3	1139	1004	1379	193	186	162	29	30

4	1520	1268	1110	142	104	176	24	36
5	1042	1223	1174	148	148	119	21	17
6	1054	1080	1911	177	185	180	29	33
7	1458	1905	1188	185	199	140	14	22
8	1307	1020	1243	110	138	129	15	12
9	1277	1213	1349	112	106	126	13	22
10	1423	1318	1102	129	158	121	19	32
11	1365	1166	1198	143	189	135	24	27
12	1679	1358	1882	150	191	108	37	19
13	1361	1079	1031	183	109	139	27	28
14	1645	1525	1834	179	174	185	20	13
15	1211	1204	1257	187	157	125	26	36
16	1291	1226	1964	129	110	192	21	14
17	1674	1620	1506	108	111	191	34	10
18	1314	1064	1075	195	163	134	40	19
19	1424	1234	1707	108	135	127	37	38
20	1666	1909	1568	112	196	118	29	16

Практическое занятие №7 Объектно-событийное моделирование сложных систем

Задача 1

Используя средства проектирования BPwin/ Ramus/ Business studio:

1. Создайте иерархическую IDEF0-модель проекта работы по теме вашего диссертационного исследования. Окончательная модель должна содержать четыре уровня иерархии (A-0 (контекстная диаграмма), A0 (основные процессы), A1…A6 и 3 диаграммы декомпозиции 4 уровня по выбору аспиранта).

2. Для полученной модели создайте дерево функций и организационную модель.

3. Проделайте процесс слияния и расщепления моделей.

4. Проведите количественный анализ полученной модели (рассчитать коэффициент декомпозиции и сбалансированности).

Разработайте комплексную BPWin-модель, состоящую из трех видов диаграмм: IDEF0, DFD и IDEF3. Контекстная диаграмма уровня A-0 и диаграмма уровня A0, с использованием IDEF0-методологии, затем 3 блока декомпозируются на DFD-диаграммы и по 1 блоку каждого уровня DFD декомпозируются на IDEF3 (3 IDEF3-диаграммы). Таким образом, должна получиться модель, состоящая из 8 диаграмм.

Практическое занятие № 8 Метод Монте-Карло

Согласно варианту задания разработать алгоритм для решения задачи на основе метода Монте-Карло. Выполнить три испытания разработанного алгоритма.

Вариант 1

Предприятие имеет 20 станков типа Ст1 и 30 станков типа Ст2. На каждом из них можно выпускать детали трёх видов. Производительность станков (деталей в день) приведена в таблице 1.

Таблица 1

Станок	Деталь		
	Д1	Д2	Д3
Ст1	20	35	15
Ст2	15	30	45

Каждый станок настраивается на выпуск детали только какого-либо одного вида. В течение каждого рабочего дня предприятие должно выпускать не менее 150 деталей Д1 и не менее 100 деталей Д2. Прибыль от продажи одной детали Д1 составляет 6 ден. ед., от продажи детали Д2 – 4 ден. ед., от продажи Д3 – 8 ден. ед.

Требуется составить план использования станков, обеспечивающий получение максимальной прибыли от выпуска деталей.

Вариант 2

Фирма, владеющая двумя предприятиями, имеет возможность выполнить четыре заказа. Каждый заказ должен выполняться только одним предприятием (совместная работа предприятий над одним заказом невозможна). Каждое из предприятий может выполнять несколько заказов. Характеристики заказов приведены в таблице 2.

Таблица 2

Характеристики заказа	Заказ			
	Р1	Р2	Р3	Р4
Затраты 1-го предприятия, млн. ден. ед.	4	2	6	3
Затраты 2-го предприятия, млн. ден. ед.	7	1	2	5
Плата заказчика, млн. ден. ед.	12	4	8	6

Это означает, например, что заказчик платит фирме за выполнение заказа Р1 12 млн. ден. ед. Если этот заказ будет выполняться первым предприятием, то затраты на его выполнение составят 4 млн. ден. ед.; таким образом, прибыль фирмы составит 12-4=8 млн. ден. ед. Если этот заказ будет выполняться вторым предприятием, то затраты составят 7 млн. ден. ед. (прибыль фирмы составит 12-7=5 млн. ден. ед).

Требуется распределить заказы между предприятиями таким образом, чтобы прибыль фирмы была максимальной.

Вариант 3

В цехе имеются два станка (СТ1 и СТ2). Каждый из станков может выпускать детали трех типов (Д1, Д2, Д3). Производительность станков (деталей в день) приведена в таблице 3.

Таблица 3

Станок	Деталь		
	Д1	Д2	Д3
СТ1	80	100	70
СТ2	100	70	50

В начале каждого рабочего дня станок налаживается на выпуск деталей некоторого типа и выпускает их весь день. В течение рабочей недели (5 дней) недели предприятию необходимо выпустить не менее 300 деталей Д1 и не менее 200 деталей Д2. Прибыль предприятия от продажи детали Д1 составляет 3 ден. ед., Д2 – 5 ден. ед., Д3 – 8 ден. ед. Прибыль не зависит от того, на каком станке выпущена деталь.

Требуется составить оптимальный план использования станков в течение рабочей недели, т.е. определить, сколько дней каждый станок должен использоваться для выпуска деталей каждого типа, чтобы общая прибыль от выпущенных деталей была максимальной.

Вариант 4

Для обслуживания трёх местных авиалиний в течение года могут использоваться самолеты двух типов: DC-3 и A-28. Имеется пять самолетов DC-3 и восемь самолетов A-28. В течение года каждый самолет совершает ровно 100 рейсов. За один рейс самолет DC-3 перевозит 30 пассажиров, A-28 – 40 пассажиров. Прибыль от одного рейса самолёта каждого типа по каждой авиалинии (тыс. ден. ед.) приведена в таблице 4.

Таблица 4

Тип самолёта	Авиалиния		
	L1	L2	L3
DC-3	2	1,5	1,2
A-28	2,2	2	1,5

Каждый самолёт на весь год закрепляется за какой-либо одной авиалинией. В течение года требуется перевезти по линии L1 не менее 12 тыс. пассажиров, по линии L3 – не менее 10 тыс.

Требуется распределить самолеты по авиалиниям на год таким образом, чтобы прибыль от авиаперевозок была максимальной.

Вариант 5

Для выращивания зерновых культур (озимых и яровых) может использоваться 0,8 млн. га земли в климатической зоне А, и 0,6 млн. га, в климатической зоне В. Урожайность земель приведена в таблице 5.

Таблица 5

Зерновые культуры	Урожайность, ц/га	
	зона А	зона В
Озимые	20	25
Яровые	24	15

Необходимо получить не менее 20 млн. центнеров озимых и не менее 6 млн. центнеров яровых. Прибыль от продажи одного центнера озимых составляет 10 ден. ед., одного центнера яровых – 8 ден. ед.

Требуется составить план использования земель, обеспечивающий максимальную прибыль от продажи урожая.

Вариант 6

Имеется 10 единиц груза ГР1, 8 единиц груза ГР2, 5 единиц груза ГР3, 10 единиц груза ГР4. Вес единицы груза следующий: ГР1 – 2 т, ГР2 – 4 т, ГР3 – 6 т, ГР4 – 1,5 т. Имеется транспортное средство грузоподъёмностью 80 т.

Требуется определить, сколько единиц груза каждого типа требуется поместить в транспортное средство, чтобы перевезти груз максимального веса.

ВОПРОСЫ К ЗАЧЕТУ.

1. Статистическое наблюдение: понятие, сущность, содержание.
2. Статистическая обработка экспериментальных данных: назначение, методы, обработка результатов
3. Оценка статистических параметров: параметры, критерии оценки.
4. Методы обработки и анализа статистической информации.
5. Средние величины и вариации.
6. Выборзаочное наблюдение: виды выборки, способы отбора, ошибки, методы.
7. Парная линейная регрессия: условия и порядок построения, анализ и направления использования.
8. Статистическое распределение: понятие, виды, принципы применения.
9. Корреляционно-регрессионный анализ: условия применения, сущность, алгоритм, интерпретация результатов.
10. Факторный анализ данных: условия применения, сущность, алгоритм, интерпретация результатов.
11. Кластерный анализ: условия применения, сущность, алгоритм, интерпретация результатов.
12. Дискриминантный анализ: условия применения, сущность, алгоритм, интерпретация результатов.
13. Робастное оценивание: условия применения, сущность, алгоритм, интерпретация результатов.
14. Интуитивные модели прогнозирования: понятие, методика построения, применение.
15. Модели временных рядов: понятие, условия применения, алгоритм построения.
16. Экспоненциальное сглаживание: понятие, назначение, алгоритм.
17. Полиномиальные модели: понятие, условия применения, алгоритм построения.
18. Методы выделения сезонных и циклических колебаний, адаптивные методы прогнозирования, прогнозирование на основе индикаторов.
19. Оценка адекватности и точности прогнозов.
20. Когнитивное моделирование: понятие, алгоритм построения.
21. Генетические методы прогнозирования: понятие, особенности, применение.
22. Динамические модели: понятие, условия применения, алгоритм построения.
23. Объектно-событийное моделирование сложных систем.
24. Проблемно-ориентированные интерактивные системы.
25. Структурные модели: понятие, условия применения, алгоритм построения.
26. Линейное и нелинейное программирование: принципы построения, применение.

Литература

Основная

1. Аверченков В.И. Основы математического моделирования технических систем/ В.И. Аверченков, В.П. Федоров, М.Л. Хейфец — Москва: Флинта, 2011 г.— 271 с. — Электронное издание. Режим доступа: http://ibooks.ru
2. Волкова, В. Н. Теория систем и системный анализ / В. Н. Волкова, А. А. Денисов. - М. : Юрайт, 2012. - 678,
3. Елисеева, И. И. Статистика/ И. И. Елисеевой ; С.-Петерб. гос. ун-т экономики и финансов. - 3-е изд., перераб. и доп. - М. : Юрайт, 2012. - 558 с.
4. Кремер, Н. Ш. Эконометрика / Н. Ш. Кремер, Б. А. Путко ; под ред. Н. Ш. Кремера. - 3-е изд., перераб. и доп. - Москва : ЮНИТИ, 2010. - 328 с.

Дополнительная

1. Белоусова, Е. Н. Практикум по основам научных исследований в агрономии/ Е. Н. Белоусова, А. А. Белоусов ; М-во сел. хоз-ва Рос. Федерации; Краснояр. гос. аграр. ун-т. - Красноярск : КрасГАУ, 2010. - 203 с.
2. Валентинов, В. А. Эконометрика: практикум / В. А. Валентинов. - 3-е изд. - Москва : Дашков и К, 2010. - 435 с.
3. Воскобойников, Ю. Е. Регрессионный анализ данных в пакете MATHCAD [Текст] : учебное пособие / Ю. Е. Воскобойников. - СПб. ; М. ; Краснодар : Лань, 2011. - 223 с
4. Бутакова, М. М. Экономическое прогнозирование: методы и приемы практических расчетов / М. М. Бутакова. - 2-е изд., испр. - М. : КноРус, 2010
5. Кремер, Н. Ш. Теория вероятностей и математическая статистика [Текст] : учебник для вузов / Н. Ш. Кремер. - М. : ЮНИТИ-ДАНА, 2000. - 542, [1] с.
6. Гмурман, В. Е. Теория вероятностей и математическая статистика: [учебное пособие для студентов вузов] / В. Е. Гмурман. - 7-е изд., стер. - М. : Высшая школа, 1999. - 478с.