Министерство образования и науки Российской Федерации

Государственное образовательное учреждение
высшего профессионального образования
«Оренбургский государственный университет»

Кафедра математических методов и моделей в экономике

А.Г. Реннер, О.С. Чудинова

ПАРАМЕТРИЧЕСКИЙ ДИСКРИМИНАНТНЫЙ АНАЛИЗ В ПАКЕТАХ STATISTICA, STATA, EXCEL

Методические указания
к лабораторному практикуму, курсовой работе, дипломному проектированию и самостоятельной работе студентов

Рекомендовано к изданию Редакционно-издательским советом Государственного образовательного учреждения высшего профессионального образования «Оренбургский государственный университет»

Оренбург
ИПК ГОУ ОГУ
2010

УДК 519.237:004.42 (07)
ББК 22.172+32.973-018.2 я73
Р 39

Рецензент
кандидат экономических наук, доцент С.В. Дьяконова

Реннер, А.Г.
Р 39 Параметрический дискриминантный анализ в пакетах Statistica, Stata, Excel: методические указания к лабораторному практикуму, курсовой работе, дипломному проектированию и самостоятельной работе студентов / А.Г. Реннер, О.С. Чудинова; Оренбургский гос. ун-т.– Оренбург: ОГУ, 2010. – 50 с.

Методические указания к семинарским занятиям, к лабораторному практикуму, самостоятельной работе студентов, в том числе для выполнения РГЗ, курсовых и дипломных работ, связанных с анализом многомерных статистических данных. Предназначены для специальности 080116 – «Математические методы в экономике», направлений 231300 – «Прикладная математика», 080500 – «Бизнес-информатика» и других специальностей и направлений, изучающих дисциплины, связанные с математическим анализом многомерных статистических данных.

УДК 519.237:004.42 (07)
ББК 22.172+32.973-018.2 я73

Содержание

Введение

Методические указания посвящены дискриминантному анализу, предназначенному для разделения рассматриваемой совокупности объектов или явлений на заданные обучающими выборками классы. Под классом понимается генеральная совокупность, заданная одномодальной функцией плотности распределения (или одномодальным полигоном вероятностей), которая в параметрическом случае считается известной с точностью до параметров. В основе классификации лежит оптимальная (байесовская) процедура отнесения объекта к тому или иному классу с минимальными потерями по сравнению с другими процедурами классификации.

В теоретической части предлагаемых методических указаниях изложены основные теоретические вопросы дискриминантного анализа. Подробно рассмотрена процедура классификации в случае нормального закона распределения классов, чаще всего применяемая на практике и реализуемая в статистических пакетах. Приведен широкий перечень теоретических вопросов и заданий по теме «Дискриминантный анализ», позволяющий студенту систематизировать свои знания и облегчить подготовку к семинарскому занятию. Часть вопросов в достаточном объеме освещены в методических указаниях и приведенной литературе, для ряда вопросов и заданий даются ссылки на необходимые источники. В практической части на конкретном примере описывается алгоритм реализации параметрического дискриминантного анализа в статистическом пакете Statistica и надстройке AtteStat пакета Excel, приводится интерпретация полученных результатов классификации. В методических указаниях сформулирована постановка задачи и определены варианты заданий, приведены требования к оформлению отчета и вопросы к защите лабораторной работы.

Использование предлагаемых методических указаний в учебном процессе позволит студенту в достаточной степени овладеть методом классификации объектов при наличии обучающей информации о классах и приобрести навыки его практической реализации в пакетах прикладных программ.

1 Теоретическая часть

1.1 Постановка задачи классификации в дискриминантном анализе

Ставится задача отнести каждый из *n* объектов, подлежащих классификации, к одному из *p* классов. Дадим определение класса: под классом в дискриминантном анализе понимается генеральная совокупность, описываемая одномодальной функцией плотности (или одномодальным полигоном вероятностей в случае дискретных признаков).

Исходная информация для анализа состоит из двух частей:

1) матрица типа «объект-свойство», содержащая информацию о значениях признаков $x_1, x_2, ..., x_k$ для *n* объектов, подлежащих классификации

$$\underset{n\times k}{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix},$$

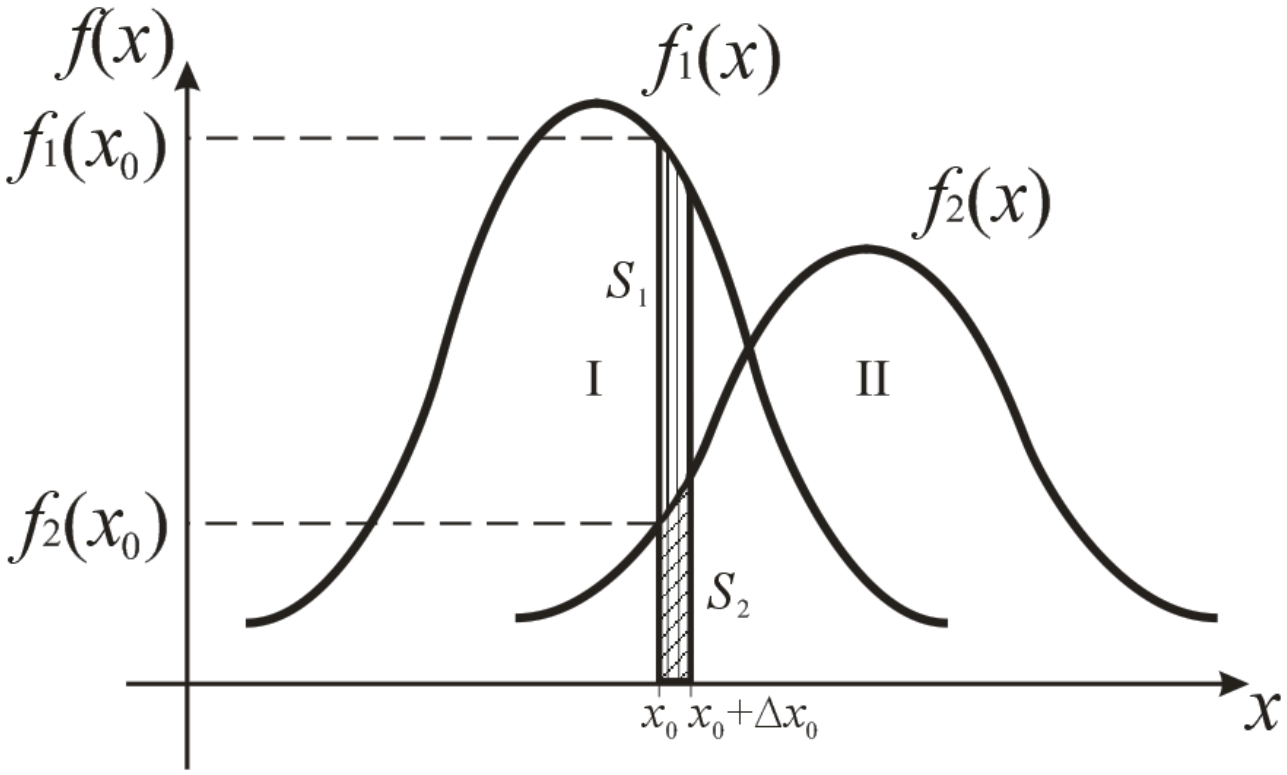
где x_{ij} - наблюдаемое значение признака x_j для *i*-го объекта выборочной совокупности, $i = \overline{1,n}$, $j = \overline{1,k}$;

2) обучающие выборки $O_1^{(j)}, O_2^{(j)}, ..., O_{n_j}^{(j)}$, $j = \overline{1,p}$. Относительно объектов $O_1^{(j)}, O_2^{(j)}, ..., O_{n_j}^{(j)}$ известно, что они принадлежат *j*-му классу и каждый из объектов характеризуется наблюдаемыми значениями *k* признаков $x_1, x_2, ..., x_k$: $O_i^{(j)} = \left(x_{i1}^{(j)}, x_{i2}^{(j)}, ..., x_{ik}^{(j)}\right)^T$, $i = \overline{1,n_j}$. Статистическую информацию по *j*-ой обучающей выборки можно представить в виде матрицы $X^{(j)}$ типа «объект-свойство»

$$\underset{n_j*k}{X}^{(j)} = \begin{pmatrix} x_{11}^{(j)} & x_{12}^{(j)} & \dots & x_{1k}^{(j)} \\ x_{21}^{(j)} & x_{22}^{(j)} & \dots & x_{2k}^{(j)} \\ \dots & \dots & \dots & \dots \\ x_{n_j1}^{(j)} & x_{n_j2}^{(j)} & \dots & x_{n_jk}^{(j)} \end{pmatrix}, \; j = \overline{1,p}.$$

Основной принцип вероятностных методов классификации заключается в следующем: объект следует отнести к тому классу (т.е. к той генеральной совокупности), в рамках которого он выглядит более правдоподобным. Иллюстрация этого принципа представлена на рисунке 1.

Сформулированный принцип может корректироваться с учетом удельных весов классов и потерь от неправильной классификации объектов.



$$P(x_0 \in \text{I}) = S_1 \approx f_1(x_0) \cdot \Delta x_0$$
$$P(x_0 \in \text{II}) = S_2 \approx f_2(x_0) \cdot \Delta x_0$$

S_1 и S_2 - площади соответствующих фигур на графике

$$S_1 > S_2 \Rightarrow \quad x_0 \in \text{I}$$

Рисунок 1 – Принцип классификации в дискриминантном анализе

1.2 Функции потерь и вероятности неправильной классификации

Очевидно, что методы классификации желательно строить так, чтобы минимизировать потери или вероятность неправильной классификации объектов.

Обозначим через $C(j/i)$ потери, которые мы несем при отнесении одного объекта *i*-ого класса к классу с номером *j*, $i, j = \overline{1, p}$. При $i = j$ $C(j/i) = 0$. Если в процессе классификации объект *i*-ого класса будет отнесен к классу с номером *j* $m(j/i)$ раз, то потери составят $m(j/i) \cdot C(j/i)$, а величина общих потерь тогда определяется следующим образом:

$$C_n = \sum_{i=1}^{p} \sum_{j=1}^{p} m(j/i) \cdot C(j/i). \qquad (1)$$

Для того, чтобы потери не зависели от числа *n* классифицируемых объектов (а величина C_n будет расти с ростом *n*), перейдем к удельной характеристике потерь, разделив обе части выражения (1) на *n* и перейдя к пределу по $n \to \infty$:

$$C = \lim_{n \to \infty} \left(\frac{1}{n} C_n \right) = \lim_{n \to \infty} \sum_{i=1}^{p} \sum_{j=1}^{p} C(j/i) \frac{m(j/i) \cdot n_i}{n_i \cdot n} = \sum_{i=1}^{p} \pi_i \sum_{j=1}^{p} C(j/i) P(j/i). \qquad (2)$$

Предел в выражении (2) следует понимать в смысле сходимости по вероятности величины $\frac{m(j/i)}{n_i}$ к $P(j/i)$ - вероятности отнесения объекта *i*-ого класса к классу *j* и величины $\frac{n_i}{n}$ к π_i - вероятности извлечения объекта *i*-ого класса из общей совокупности объектов. Величину π_i называют априорной вероятностью или удельным весом *i*-ого класса.

Величина $C^{(i)} = \sum_{j=1}^{p} C(j/i) P(j/i)$ определяет средние потери от неправильной классификации объектов *i*-ого класса. Тогда средние удельные потери от неправильной классификации всех анализируемых объектов составляют: $C = \sum_{i=1}^{p} \pi_i C^{(i)}$.

Часто полагают, что потери $C(j/i)$ одинаковы для любой пары *i* и *j*, т.е. $C(j/i) = C_0 = const \quad \forall i, j = \overline{1, p}$, $i \neq j$. В этом случае стремление минимизировать

средние удельные потери C будет эквивалентно стремлению максимизировать вероятность правильной классификации объектов равной $\sum_{j=1}^{p} \pi_i P(i/i)$.

1.3 Построение оптимальных (байесовских) процедур классификации

Классифицируемые наблюдения в дискриминантном анализе интерпретируются как выборка из генеральной совокупности, описываемой смесью k классов, с плотностью распределения $f(x) = \sum_{j=1}^{p} \pi_j f_j(x)$, где $f_j(x)$ - плотность распределения j-ого класса, π_j - априорная вероятность появления объекта j-ого класса или удельный вес объектов j-ого класса в общей генеральной совокупности, $j = \overline{1, p}$.

Введем понятие процедуры классификации, т.е. решающего правила отнесения объекта, характеризующегося многомерным вектором признаков $x = (x_1, x_2, ..., x_k)$, к j-ому классу. Для этого строится дискриминантная функция $\delta(x)$, принимающая только целые положительные значения 1, 2, …, p, причем те x, для которых функция принимает значение, равное j, относят к классу j, т.е. $S_j = \{x : \delta(x) = j\}$, $j = \overline{1, p}$. Таким образом получаем, что S_j - это k-мерная область в пространстве $П(x)$ возможных значений анализируемого многомерного признака x. Функция $\delta(x)$ строится таким образом, чтобы теоретико-множественная сумма $S_1 + S_2 + ... + S_p$ заполняла все пространство $П(x)$ и чтобы области S_j, $j = \overline{1, p}$ попарно не пересекались. Таким образом, решающее правило $\delta(x)$ может быть задано разбиением $S = (S_1, S_2, ..., S_p)$ всего пространства $П(x)$ на p непересекающихся областей.

Процедура классификации называется оптимальной (байесовской), если она сопровождается минимальными потерями (2) среди всех других процедур классификации. Процедура классификации $S^{опт} = \left(S_1^{опт}, S_2^{опт}, ..., S_p^{опт}\right)$, при которой потери (2) будут минимальными, определяется следующим образом:

$$S_j^{опт} = \left\{ x : \sum_{\substack{i=1 \\ i \neq j}}^{p} \pi_i f_i(x) C(j/i) = \min_{1 \le l \le k} \sum_{\substack{i=1 \\ i \neq l}}^{p} \pi_i f_i(x) C(l/i) \right\} . \quad (3)$$

Таким образом, наблюдение $x_v = (x_{v1}, x_{v2}, \ldots, x_{vk})^T$, $v = \overline{1,n}$ будет отнесено к классу j тогда и только тогда, когда средние удельные потери от его отнесения именно в этот класс окажутся минимальными по сравнению с аналогичными потерями, связанными с отнесением этого наблюдения в любой другой класс.

В случае равных потерь $C(j/i) = C_0 = const \quad \forall i, j = \overline{1,p}$, $i \neq j$ правило классификации приобретает более простой вид: объект x_v будет отнесен к классу j тогда и только тогда, когда

$$\pi_j f_j(x_v) = \max_{1 \le l \le p} \pi_l f_l(x_v) , \quad (4)$$

т.е. максимизируется «взвешенная правдоподобность» этого объекта в рамках класса, где в качестве весов выступают априорные вероятности.

Выражения (3) и (4) задают теоретическое оптимальное правило классификации. Для того, чтобы его реализовать, необходимо знать априорные вероятности π_j и законы распределения классов $f_j(x)$, $j = \overline{1,p}$. В статистическом варианте решения этой задачи перечисленные характеристики заменяются соответствующими оценками, построенными на базе обучающих выборок.

Если данные, составленные из всех обучающих выборок, можно считать случайной выборкой объемом $n_{об} = n_1 + n_2 + \ldots + n_p$, то оценки удельных весов классов π_j, $j = \overline{1,p}$, можно рассчитать по формуле: $\hat{\pi}_j = \frac{n_j}{n_{об}}$.

Что касается задачи оценки законов распределения $f_j(x)$, $j = \overline{1,p}$, то ее удобно разбить на два случая:

1) параметрический дискриминантный анализ: вид функций $f_j(x)$, $j = \overline{1,p}$, известен, не известны параметры распределения классов. В качестве оценки $f_j(x)$

выступает $f_j\left(x, \hat{\theta}^{(j)}\right)$, где $\hat{\theta}^{(j)}$ – оценка параметров распределения j-го класса, рассчитанные на основе j-ой обучающей выборки;

2) непараметрический дискриминантный анализ: вид функций $f_j(x)$, $j = \overline{1, p}$, не известен. В этом случае строят непараметрические оценки функций $f_j(x)$, например, гистограммного или ядерного типа, либо пользуются некоторыми специальными приемами.

1.4 Параметрический дискриминантный анализ в случае нормального закона распределения классов

Пусть класс j, $j = \overline{1, p}$, идентифицируется как k-мерная нормально распределенная генеральная совокупность с вектором математических ожиданий $a^{(j)} = \left(a_1^{(j)}, a_2^{(j)}, ..., a_k^{(j)}\right)^T$ и ковариационной матрицей Σ общей для всех классов.

Перепишем правило классификации (4) следующим образом: объект x_v относится к классу j тогда и только тогда, когда

$$\frac{f_j(x_v)}{f_l(x_v)} \geq \frac{\pi_l}{\pi_j} \quad \forall l = \overline{1, p}. \tag{5}$$

Прологарифмируем левую и правую часть выражения (5):

$$\ln\left(\frac{f_j(x_v)}{f_l(x_v)}\right) \geq \ln\left(\frac{\pi_l}{\pi_j}\right) \quad \forall l = \overline{1, p}. \tag{6}$$

В случае нормального закона распределения классов плотность распределения $f_l(x)$, $l = \overline{1, p}$, имеет вид:

$$f_l(x) = \frac{1}{(2\pi)^{k/2}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}\left(x - a^{(l)}\right)^T \Sigma^{-1}\left(x - a^{(l)}\right)\right). \tag{7}$$

Подставим (7) в выражение (6) и после ряда преобразований получим правило классификации в случае нормального закона распределения классов с равными ковариационными матрицами. Оно формулируется следующим образом: объект x_v относится к классу j тогда и только тогда, когда

$$\left[x_v - \frac{1}{2}\left(a^{(j)} + a^{(l)}\right)\right]^T \Sigma^{-1}\left(a^{(j)} - a^{(l)}\right) \geq \ln\frac{\pi_l}{\pi_j} \quad \forall l = \overline{1, p}. \tag{8}$$

Для реализации правила классификации (8) необходимо знать параметры распределения классов $a^{(j)} = \left(a_1^{(j)}, a_2^{(j)}, ..., a_k^{(j)}\right)^T$, Σ и удельные веса классов π_j, $j = \overline{1, p}$. Если перечисленные характеристики не известны, то на основе обучающих выборок рассчитываются их оценки $\hat{a}^{(j)} = \bar{x}^{(j)} = \left(\bar{x}_1^{(j)}, \bar{x}_2^{(j)}, ..., \bar{x}_k^{(j)}\right)^T$, $\hat{\Sigma}$, $\hat{\pi}_j$, где $\bar{x}_s^{(j)}$ - среднее арифметическое значение признака x_s, рассчитанное на основе j-ой обучающей выборки. Оценка ковариационной матрицы, общей для всех классов, рассчитывается по формуле

$$\hat{\Sigma} = \frac{1}{n_{об} - p}\left[(n_1 - 1)\hat{\Sigma}^{(1)} + ... + (n_p - 1)\hat{\Sigma}^{(p)}\right], \tag{9}$$

где $\hat{\Sigma}^{(j)}$ – оценка ковариационной матрицы, рассчитанная на основе j-ой обучающей выборки.

Таким образом, правило классификации (9) в выборочном случае имеет вид: объект x_v относится к классу j тогда и только тогда, когда

$$\left[x_v - \frac{1}{2}\left(\hat{a}^{(j)} + \hat{a}^{(l)}\right)\right]^T \hat{\Sigma}^{-1}\left(\hat{a}^{(j)} - \hat{a}^{(l)}\right) \geq \ln\frac{\hat{\pi}_l}{\hat{\pi}_j} \quad \forall l = \overline{1, p}. \tag{10}$$

Правило (10) можно преобразовать к виду:

$$x_v^T \hat{\Sigma}^{-1} \hat{a}^{(j)} - \frac{1}{2} \hat{a}^{(j)T} \hat{\Sigma}^{-1} \hat{a}^{(j)} + \ln \hat{\pi}_j \geq x_v^T \hat{\Sigma}^{-1} \hat{a}^{(l)} - \frac{1}{2} \hat{a}^{(l)T} \hat{\Sigma}^{-1} \hat{a}^{(l)} + \ln \hat{\pi}_l \quad \forall l = \overline{1, p}.$$

Каждому классу l ставится в соответствие линейная дискриминантная функция $\varphi_l(x) = b_0^{(l)} + b_1^{(l)} x_1 + b_2^{(l)} x_2 + \ldots + b_k^{(l)} x_k = b_0^{(l)} + b^{(l)} x$, $\forall l = \overline{1, p}$, где $b_0^{(l)} = -\frac{1}{2} \hat{a}^{(l)T} \hat{\Sigma}^{-1} \hat{a}^{(l)} + \ln \hat{\pi}_l$, $b^{(l)} = \left(b_1^{(l)}, b_2^{(l)}, \ldots, b_k^{(l)}\right) = \hat{\Sigma}^{-1} \hat{a}^{(l)}$. Тогда объект x_v относится к классу *j* тогда и только тогда, когда

$$\varphi_j(x_v) = \max_{1 \leq l \leq p} \varphi_l(x_v). \tag{11}$$

1.5 Геометрическая интерпретация дискриминантного анализа в случае нормального закона распределения классов

Пусть $k = 2$, $p = 2$, $\pi_1 = \pi_2$, $\Sigma = E$. Тогда объект x_0 относится к первому классу если:

$$\left[x_0 - \frac{1}{2} \left(\hat{a}^{(1)} + \hat{a}^{(2)} \right) \right]^T \left(\hat{a}^{(1)} - \hat{a}^{(2)} \right) \geq 0. \tag{12}$$

Геометрическая интерпретацию правила (12) представлена на рисунке 2.

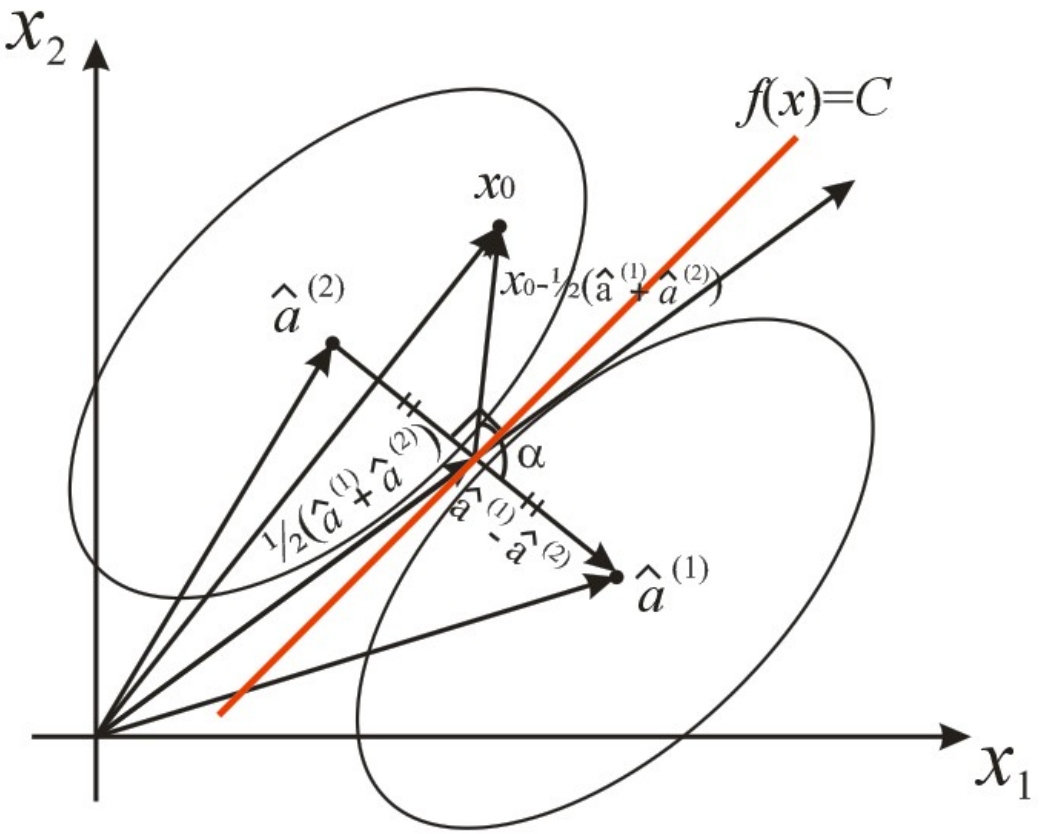


Рисунок 2 – Геометрическая интерпретация дискриминантного анализа в двумерном случае

Знак в левой части неравенства (12) зависит от угла α . Если угол α , как в нашем случае, тупой, то $\cos\alpha < 0$, следовательно, объект x_0 следует отнести ко второму классу. Таким образом, все объекты, лежащие слева от прямой, перпендикулярной вектору $\hat{a}^{(1)} - \hat{a}^{(2)}$ и проходящей через его середину, относятся ко второму классу, а все объекты, лежащие справа от прямой, относятся к первому классу. Прямая $f(x) = b_1x_1 + b_2x_2 = C$ наилучшим образом разделяет два класса объектов и называется дискриминантной прямой, константа C называется константой дискриминации.

1.6 Вопросы и задания, выносимые на семинарские занятия, по теме «Дискриминантный анализ»

1) Что понимается под классификацией в дискриминантном анализе?
2) Дайте определение классификации с «обучением»

3) Что называется обучающей выборкой?
4) Сформулируйте постановку задачи классификации в дискриминантном анализе
5) Что понимается под классом в дискриминантном анализе?
6) В чем отличие параметрического и непараметрического дискриминантного анализа?
7) Сформулируйте основной принцип вероятностных методов классификации и проиллюстрируйте его на графике
8) Выведите формулу для определения средних удельных потерь от неправильной классификации
9) Докажите, что задача минимизации средних удельных потерь эквивалентна задаче максимизации вероятности правильной классификации [5, с. 472-473]
10) Используя формулу полной вероятности, выведите плотность распределения смеси *p* классов
11) Что называется процедурой классификации?
12) Какая процедура классификации называется оптимальной (байесовской)?
13) Сформулируйте правило классификации объектов в случае постоянных потерь от неправильной классификации. Как пользоваться этим правилом на практике?
14) Приведите примеры законов распределения классов, для которых применимо правило классификации (4). Опишите алгоритм реализации этого правила в каждом случае
15) Выведите правило классификации объектов в случае нормального закона распределения классов с равными ковариационными матрицами
16) Запишите правило классификации (8) на следующие случаи:
    - объект x_v относится к первому классу;
    - объект x_v относится ко второму классу, количество классов $p = 2$;
    - удельные веса классов одинаковые;
    - количество признаков $k = 1$, количество классов $p = 2$, $\pi_1 = \pi_2$ [5, с. 477].
17) Как на практике реализовать правило классификации, полученное в задании 15? Запишите формулы для расчета оценок $a^{(j)}$ и $\Sigma^{(j)}$ - параметров нормально распределенного *j*-го класса [5, с. 476], [2, с. 257]

18) Сформулируйте правило классификации, полученное в задании 15, через линейные дискриминантные функции
19) Сформулируйте условия использования каждого из правил классификации (4), (8), (11)
20) Дайте геометрическую интерпретацию дискриминантного анализа в случае нормального закона распределения классов
21) Каким образом классифицировать объект x_v, лежащий на дискриминантной прямой?
22) Выведите уравнение дискриминантной прямой в случае двух признаков, двух нормально распределенных классов с единичными ковариационными матрицами и одинаковыми удельными весами классов [7, с. 509-513]
23) Установите связь между правилом классификации (12) и коэффициентами в уравнении дискриминантной прямой
24) Запишите правило классификации (12) через дискриминантную функцию $f(x) = b_1x_1 + b_2x_2$ и константу дискриминации C [2, с. 259-260]
25) Каким образом зависит константа дискриминации C от удельных весов классов и как изменится положение дискриминантной прямой на рисунке 2, если $\pi_1 > \pi_2$, $\pi_1 < \pi_2$?
26) Обобщите результаты, полученные в заданиях 22-24, на k-мерный случай, отказываясь от условий $\pi_1 = \pi_2$ и $\Sigma = E$

2 Практическая часть

2.1 Содержание лабораторной работы

Выполнение лабораторной работы по теме «Дискриминантный анализ» состоит из следующих этапов:

- ознакомление с формулировкой задания к лабораторной работе и порядком её выполнения в пакетах прикладных программ;
- выполнение расчетов на компьютере по данным своего варианта;
- анализ полученных результатов;

– подготовка письменного отчета по лабораторной работе;

– защита лабораторной работы.

2.2 Задание к лабораторной работе

Районы Оренбургской области характеризуются социально-экономическими показателями, обозначение и наименование которых представлены в таблице А.1. Значения показателей для 35 районов области за 2007 год приведены в таблице А.2 [4]. Имеются p обучающих выборок из нормально распределенных генеральных совокупностей с равными ковариационными матрицами. В таблице А.3 для каждого варианта приведены набор из пяти показателей для анализа, количество и состав обучающих выборок. Ставится задача провести классификацию районов Оренбургской области, не вошедших в обучающие выборки, на p классов и дать экономическую интерпретацию результатов классификации.

2.3 Порядок выполнения лабораторной работы в пакете Statistica

Порядок выполнения лабораторной работы рассмотрен на основании данных нулевого варианта, включающего следующие показатели для анализа:

x_4 – инвестиции, направленные в жилищное хозяйство, на душу населения, руб.;

x_6 – ввод в действие жилых домов на 1000 человек населения, кв.м;

x_7 – ввод в действие жилых домов, построенных населением за свой счет и с помощью кредитов, кв.м;

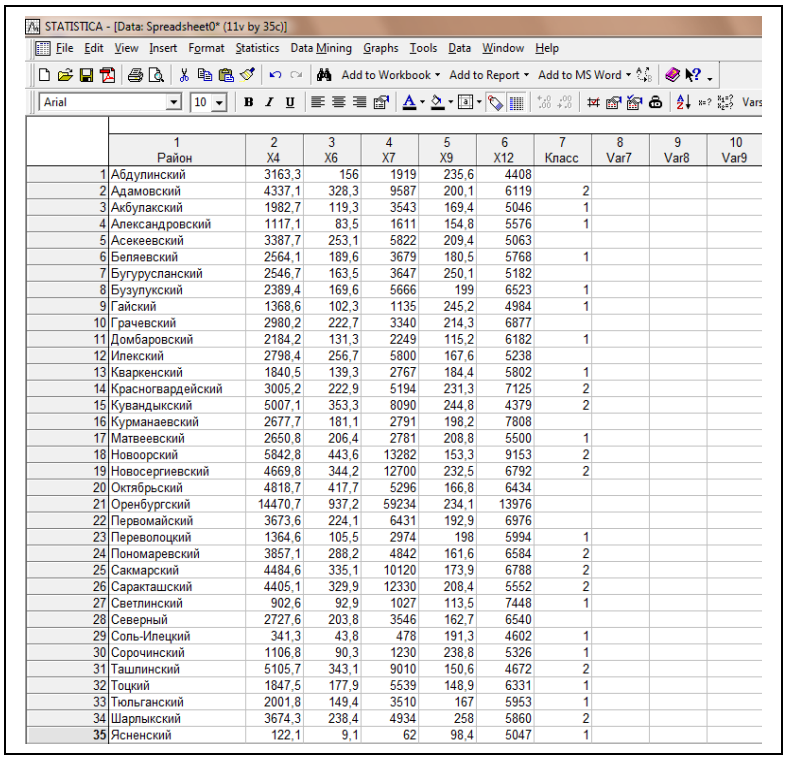
x_9 – обеспеченность населения собственными легковыми автомобилями в расчете на 1000 населения, штук;

x_{12} – среднемесячная начисленная заработная плата работников, руб.

Так как имеются обучающие выборки и известен вид закона распределения классов, то классификацию районов можно провести с помощью параметрического дискриминантного анализа. При этом необходимо проверить, чтобы число объектов

в каждой обучающей выборки было хотя бы на 2 единицы больше чем число признаков.

Вид таблицы с исходными данными для анализа в пакете Statistica 8.0 представлен на рисунке 3. В первом столбце для удобства введены названия районов, в следующих пяти столбцах введены значения социально-экономических показателей для соответствующих районов, в седьмом столбце – значения признака, указывающего на принадлежность к классу. Так для районов, относящихся по условию к первой обучающей выборке, в седьмом столбце введена цифра 1, для районов, относящихся ко второй обучающей выборке – цифра 2. Для районов, подлежащих классификации, значение признака **Класс** не указывается.


STATISTICA - [Data: Spreadsheet0* (11v by 35c)]

File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window Help

Add to Workbook Add to Report Add to MS Word

Arial 10

	1 Район	2 X4	3 X6	4 X7	5 X9	6 X12	7 Класс	8 Var7	9 Var8	10 Var9
1	Абдулинский	3163,3	156	1919	235,6	4408				
2	Адамовский	4337,1	328,3	9587	200,1	6119	2			
3	Акбулакский	1982,7	119,3	3543	169,4	5046	1			
4	Александровский	1117,1	83,5	1611	154,8	5576	1			
5	Асекеевский	3387,7	253,1	5822	209,4	5063				
6	Беляевский	2564,1	189,6	3679	180,5	5768	1			
7	Бугурусланский	2546,7	163,5	3647	250,1	5182				
8	Бузулукский	2389,4	169,6	5666	199	6523	1			
9	Гайский	1368,6	102,3	1135	245,2	4984	1			
10	Грачевский	2980,2	222,7	3340	214,3	6877				
11	Домбаровский	2184,2	131,3	2249	115,2	6182	1			
12	Илекский	2798,4	256,7	5800	167,6	5238				
13	Кваркенский	1840,5	139,3	2767	184,4	5802	1			
14	Красногвардейский	3005,2	222,9	5194	231,3	7125	2			
15	Кувандыкский	5007,1	353,3	8090	244,8	4379	2			
16	Курманаевский	2677,7	181,1	2791	198,2	7808				
17	Матвеевский	2650,8	206,4	2781	208,8	5500	1			
18	Новоорский	5842,8	443,6	13282	153,3	9153	2			
19	Новосергиевский	4669,8	344,2	12700	232,5	6792	2			
20	Октябрьский	4818,7	417,7	5296	166,8	6434				
21	Оренбургский	14470,7	937,2	59234	234,1	13976				
22	Первомайский	3673,6	224,1	6431	192,9	6976				
23	Переволоцкий	1364,6	105,5	2974	198	5994	1			
24	Пономаревский	3857,1	288,2	4842	161,6	6584	2			
25	Сакмарский	4484,6	335,1	10120	173,9	6788	2			
26	Саракташский	4405,1	329,9	12330	208,4	5552	2			
27	Светлинский	902,6	92,9	1027	113,5	7448	1			
28	Северный	2727,6	203,8	3546	162,7	6540				
29	Соль-Илецкий	341,3	43,8	478	191,3	4602	1			
30	Сорочинский	1106,8	90,3	1230	238,8	5326	1			
31	Ташлинский	5105,7	343,1	9010	150,6	4672	2			
32	Тоцкий	1847,5	177,9	5539	148,9	6331	1			
33	Тюльганский	2001,8	149,4	3510	167	5953	1			
34	Шарлыкский	3674,3	238,4	4934	258	5860	2			
35	Ясненский	122,1	9,1	62	98,4	5047	1			

Рисунок 3 – Исходные данные для анализа

Запуск модуля дискриминантного анализа осуществляется с помощью пункта меню **Statistics (Статистика)**, подпунктов **Multivariate Exploratory Techniques (Многомерные исследовательские методы)**, **Discriminant Function Analysis (Дискриминантный анализ)**. Вид экрана представлен на рисунке 4.

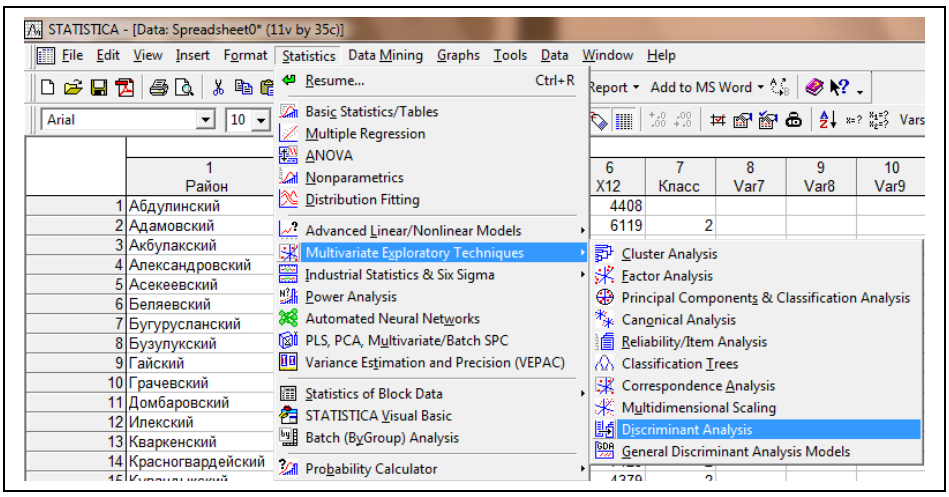


Рисунок 4 – Выбор пунктов меню

После запуска модуля дискриминантного анализа на экране появится форма, представленная на рисунке 5.

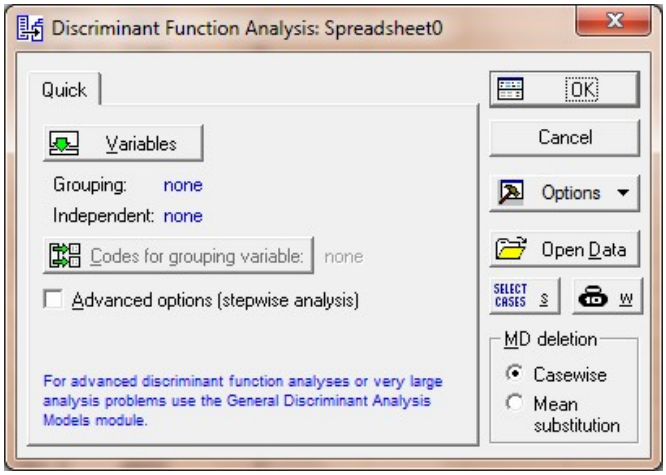


Рисунок 5 – Форма «Discriminant Function Analysis»

С помощью кнопки **Variables** необходимо выбрать признаки для анализа. Вид формы представлен на рисунке 6. В левом окне необходимо выбрать столбец матрицы исходных данных, в котором содержится номер класса (**7 - Класс**), в правом окне – столбцы, содержащие значения признаков, участвующих в анализе (**X4, X6, X7, X9, X12**).

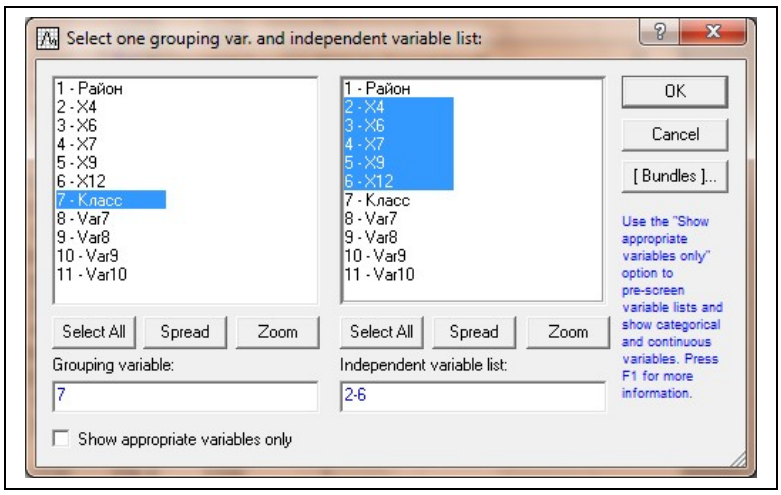


Рисунок 6 – Форма выбора признаков для дискриминантного анализа

С помощью кнопки **Codes for grouping variable** задаются коды классов (возможные значения признака **Класс**). Форма кодирования классов представлена на рисунке 7. Нажав на кнопку **All,** в поле будут автоматически введены значения **1-2**.

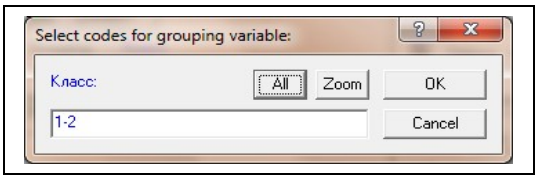


Рисунок 7 – Кодирование классов

Выбор опции **Advanced options (stepwise analysis)** на форме «Discriminant Function Analysis» позволит расширить возможности модуля, сделает доступным пошаговый отбор признаков для анализа. Вид формы представлен на рисунке 8. После нажатия на кнопку **OK** на экране появится форма выбора метода отбора признаков для анализа, представленная на рисунке 9.

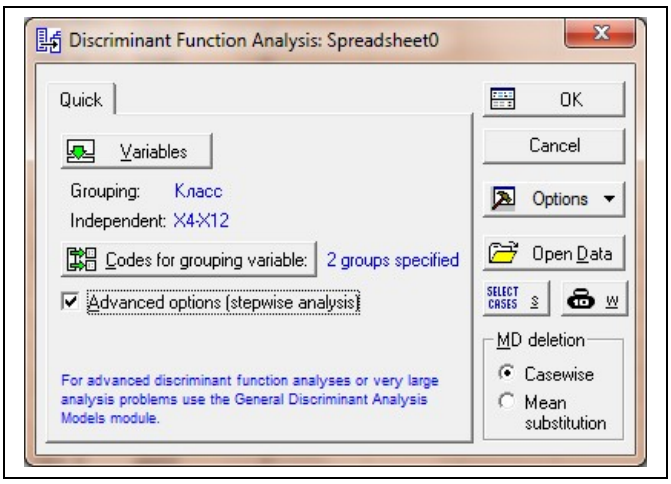


Рисунок 8 – Заполненная форма «Discriminant Function Analysis»

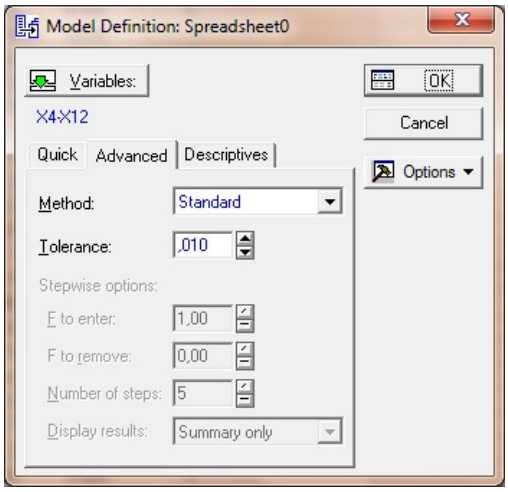


Рисунок 9 – Выбор метода отбора признаков для анализа

Метод **Standard** проводит классификацию по всем выбранным признакам. Методы **Forward** и **Backward stepwise** реализуют соответственно процедуры пошагового включения и пошагового исключения признаков, которые позволяют отобрать наиболее значимые при классификации признаки. В первом случае среди всех признаков находится тот, который вносит наибольший вклад в различие между классами. Этот признак включается в модель на первом шаге. На следующих шагах алгоритма такая процедура повторяется для оставшихся признаков. Во втором случае на первом шаге все признаки включаются в модель, а затем на каждом шаге устраняется по одному признаку, вносящему наименьший вклад в различие между классами. Пошаговые процедуры при отборе признаков «руководствуются» значениями *F*-статистики: для включения – **F to enter** и для исключения – **F to**

remove, которые задаются в диалоговом окне. Будем проводить классификацию по всем пяти признакам.

Для оценки параметров распределения в классах предназначена кнопка **Review descriptive statistics** на странице **Descriptives** формы **Model Definition**. Вид формы **Model Definition** на странице **Descriptives** представлен на рисунке 10.

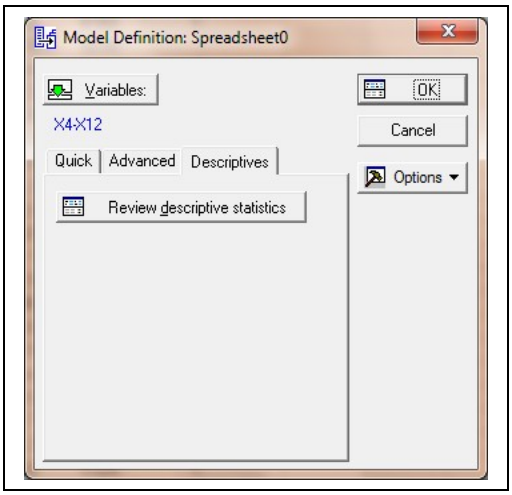


Рисунок 10 – Страница «Descriptives»

После нажатия кнопки **Review descriptive statistics** на экране появится форма, представленная на рисунке 11.

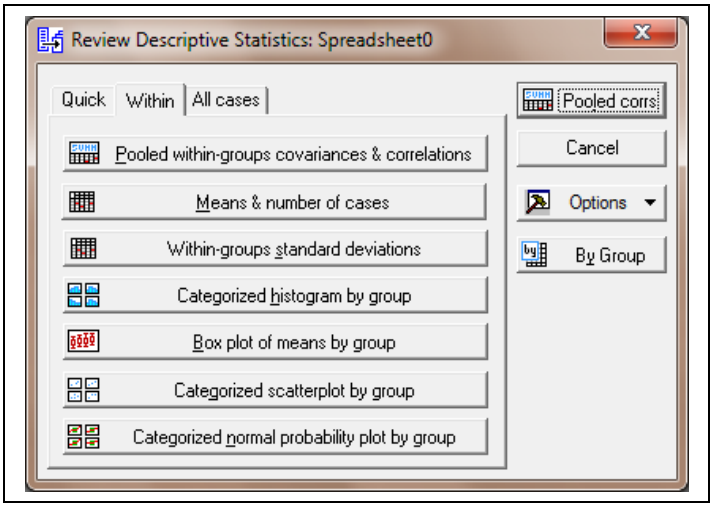
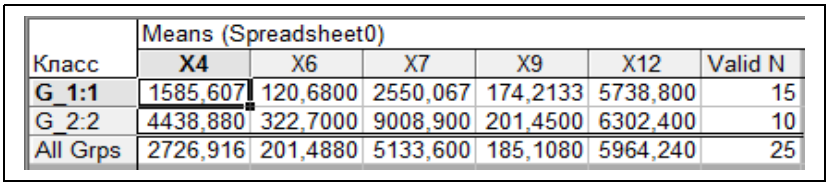


Рисунок 11 – Форма для оценок параметров распределения

Кнопка **Pooled within-groups covariances & correlations** предназначена для расчета оценок общих для двух классов ковариационной и корреляционной матриц. С помощью кнопок **Means & number of cases** и **Within-groups standard deviations** рассчитываются оценки математических ожиданий и средних квадратических отклонений признаков в классах. Средние арифметические значения признаков, рассчитанные по обучающим выборкам, представлены на рисунке 12. По полученным результатам можно дать интерпретацию классам.



	Means (Spreadsheet0)					
Класс	**X4**	X6	X7	X9	X12	Valid N
G_1:1	1585,607	120,6800	2550,067	174,2133	5738,800	15
G_2:2	4438,880	322,7000	9008,900	201,4500	6302,400	10
All Grps	2726,916	201,4880	5133,600	185,1080	5964,240	25

Рисунок 12 – Оценки математических ожиданий признаков в классах

Все средние значения показателей, рассчитанные по первой обучающей выборке, меньше соответствующих средних значений показателей, рассчитанных по второй обучающей выборке. Это позволяет сделать вывод, что по рассматриваемым показателям социально-экономическое положение районов второго класса лучше, чем первого.

Остальные кнопки на странице **Within** формы **Review Descriptive Statistics** предназначены для построения различных графиков.

После нажатия кнопки **OK** на форме **Model Definition** на экране появится форма результатов дискриминантного анализа, представленная на рисунке 13.

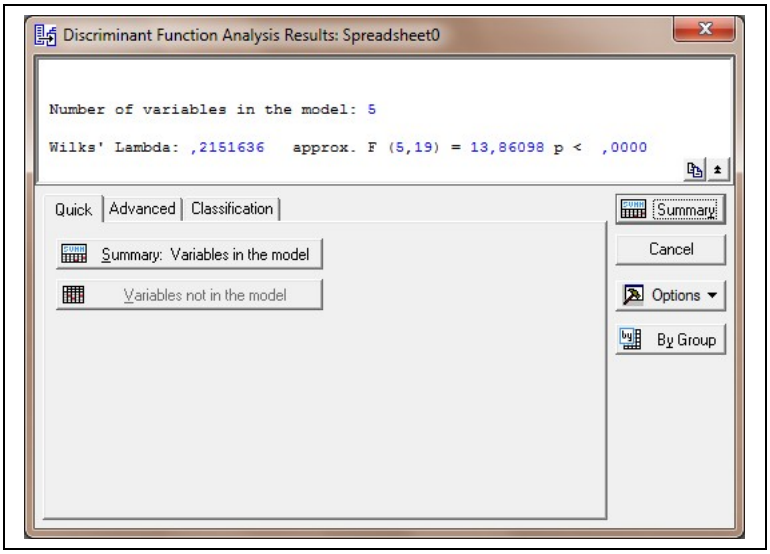


Рисунок 13 – Результаты дискриминантного анализа (страница **Quick**)

В информационной части формы представлены наблюдаемое значение статистики Уилкса, приближенное значение *F*-критерия и значимость нулевой гипотезы об отсутствии различий в групповых средних значениях всех признаков. На основе полученных результатов можно сделать вывод, что гипотеза об отсутствии различий в математических ожиданиях признаков в двух классах отвергается. Проверка такой гипотезы по каждому отдельному признаку проводится с помощью кнопки **Summary: Variables in model**.

Вид формы результатов дискриминантного анализа на странице **Classification** представлен на рисунке 14. В группе радио-кнопок **A priori classification probabilities** предложены три варианта задания априорных вероятностей:

- пропорционально объемам обучающих выборок;
- равные для всех классов;
- в результате диалога с пользователем.

С помощью кнопки **Classification functions** рассчитываются коэффициенты линейных дискриминантных функций Фишера. Результаты представлены на рисунке 15.

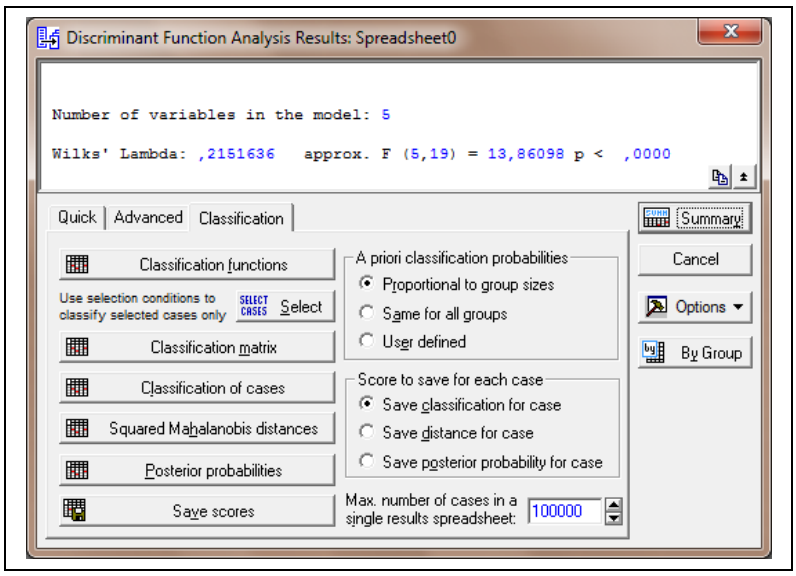


Рисунок 14 – Результаты дискриминантного анализа (страница **Classification**)

Variable	Classification Functions G_1:1 p=,60000	G_2:2 p=,40000
X4	0,01395	0,01825
X6	-0,16598	-0,16234
X7	-0,00072	-0,00066
X9	0,16716	0,18701
X12	0,00936	0,00947
Constant	-42,06674	-60,94574

Рисунок 15 – Коэффициенты линейных дискриминантных функций Фишера

В названии столбцов таблицы, представленной на рисунке 15, приведены оценки априорных вероятностей, рассчитанные по первому варианту: $\hat{\pi}_1 = 0{,}6$, $\hat{\pi}_2 = 0{,}4$. Линейные дискриминантные функции Фишера имеют вид:

$$\varphi_1(x_4, x_6, x_7, x_9, x_{12}) = -42{,}06674 + 0{,}01395x_4 - 0{,}16598x_6 - \\ - 0{,}00072x_7 + 0{,}16716x_9 + 0{,}00936x_{12}; \quad (13)$$

$$\varphi_2(x_4, x_6, x_7, x_9, x_{12}) = -60{,}94574 + 0{,}01825x_4 - 0{,}16234x_6 - \\ - 0{,}00066x_7 + 0{,}18701x_9 + 0{,}00947x_{12}. \quad (14)$$

Следует отметить, что малые значения коэффициентов в дискриминантных функциях (13), (14) связаны с большим масштабом измерения рассматриваемых показателей.

На основе функций (13), (14) проводится повторная классификация объектов обучающих выборок. Чтобы увидеть результаты этой процедуры, необходимо выбрать кнопку **Classification matrix**. На экране появится таблица, представленная на рисунке 16.

Classification Matrix (Spreadsheet0)
Rows: Observed classifications
Columns: Predicted classifications

Group	Percent Correct	G_1:1 p=,60000	G_2:2 p=,40000
G_1:1	100,0000	15	0
G_2:2	100,0000	0	10
Total	100,0000	15	10

Рисунок 16 – Результаты классификации объектов обучающих выборок

Как видно из рисунка 16, изменений в первоначальном составе классов не произошло: к первому классу относятся те же 15 районов, ко второму – те же 10 районов. Качество распознавания составило 100%. Это свидетельствует о хорошей дискриминации объектов обучающих выборок на основе функций (13), (14).

Для представления результатов классификации с помощью дискриминантных функций Фишера предназначены кнопки **Classification of cases**, **Squared Mahalanobis distances** и **Posterior probabilities**. Если объект, априори относившийся к одному классу, после реализации процедуры классификации отнесен к другому, то соответствующая этому объекту строка помечается «звездочкой» (в рассматриваемом примере такие объекты не встречаются).

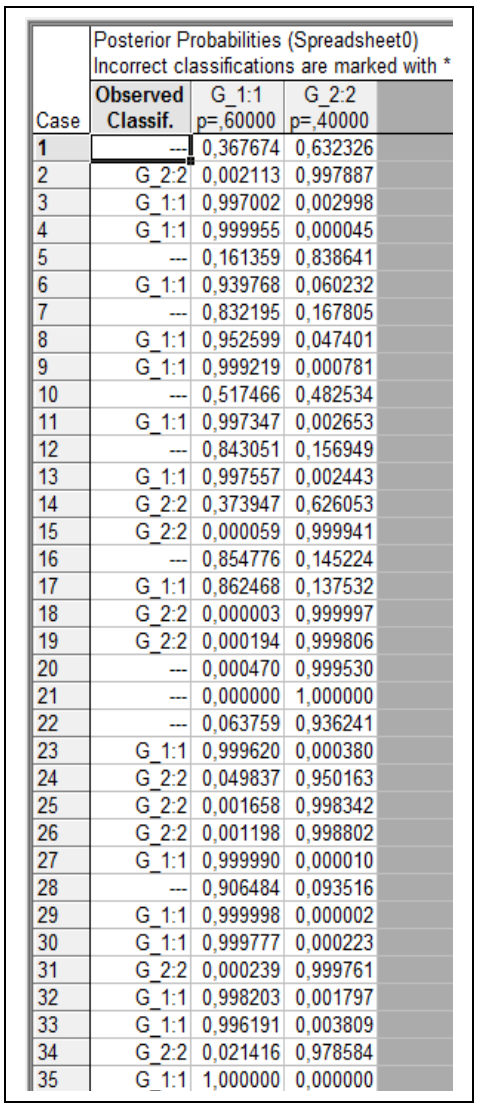
Наиболее удобны для интерпретации результаты классификации, выводимые на экран с помощью кнопок **Squared Mahalanobis distances** и **Posterior probabilities**. В первом случае рассчитываются квадраты расстояния Махаланобиса от объектов до центров каждого из классов. Результаты представлены на рисунке 17. Объект следует отнести к тому классу, расстояние до которого наименьшее. Так, например, первый район (Абдулинский) следует отнести ко второму классу,

поскольку расстояние от этого объекта до центра второго класса меньше, чем до центра первого класса (23,9204<25,8157).

Case	Squared Mahalanobis Distances from Group Centroids Incorrect classifications are marked with *		
	Observed Classif.	G_1:1 p=,60000	G_2:2 p=,40000
1	---	25,8157	23,9204
2	G_2:2	14,1035	0,9771
3	G_1:1	4,2074	15,0100
4	G_1:1	0,7181	19,9152
5	---	8,0721	3,9648
6	G_1:1	2,1184	6,8023
7	---	6,9461	9,3377
8	G_1:1	4,2925	9,4827
9	G_1:1	3,1313	16,6294
10	---	8,6464	7,9752
11	G_1:1	7,0886	18,1364
12	---	19,3138	21,8651
13	G_1:1	0,2413	11,4547
14	G_2:2	8,4641	6,6225
15	G_2:2	26,4942	6,2165
16	---	12,9224	15,6566
17	G_1:1	6,4416	9,3025
18	G_2:2	35,8292	9,6828
19	G_2:2	23,1550	5,2446
20	---	65,4449	49,3073
21	---	865,8831	752,1978
22	---	20,3537	14,1693
23	G_1:1	1,0929	16,0330
24	G_2:2	13,2487	6,5421
25	G_2:2	14,4124	0,8000
26	G_2:2	20,2459	5,9830
27	G_1:1	5,9002	28,1761
28	---	3,8072	7,5391
29	G_1:1	4,6309	29,6797
30	G_1:1	2,8258	18,8280
31	G_2:2	25,2951	7,8087
32	G_1:1	6,2521	18,0812
33	G_1:1	0,3201	10,6422
34	G_2:2	15,6940	7,2391
35	G_1:1	8,6226	39,4486

Рисунок 17 – Расстояния до центров классов

Апостериорные вероятности классификации рассчитываются с помощью кнопки **Posterior probabilities**. Результаты представлены на рисунке 18. Объект следует отнести к тому классу, апостериорная вероятность для которого наибольшая. Так, например, пятый район (Асекеевский) следует отнести ко второму классу (0,161359<0,838641).



Posterior Probabilities (Spreadsheet0)
Incorrect classifications are marked with *

Case	Observed Classif.	G_1:1 p=,60000	G_2:2 p=,40000
1	---	0,367674	0,632326
2	G_2:2	0,002113	0,997887
3	G_1:1	0,997002	0,002998
4	G_1:1	0,999955	0,000045
5	---	0,161359	0,838641
6	G_1:1	0,939768	0,060232
7	---	0,832195	0,167805
8	G_1:1	0,952599	0,047401
9	G_1:1	0,999219	0,000781
10	---	0,517466	0,482534
11	G_1:1	0,997347	0,002653
12	---	0,843051	0,156949
13	G_1:1	0,997557	0,002443
14	G_2:2	0,373947	0,626053
15	G_2:2	0,000059	0,999941
16	---	0,854776	0,145224
17	G_1:1	0,862468	0,137532
18	G_2:2	0,000003	0,999997
19	G_2:2	0,000194	0,999806
20	---	0,000470	0,999530
21	---	0,000000	1,000000
22	---	0,063759	0,936241
23	G_1:1	0,999620	0,000380
24	G_2:2	0,049837	0,950163
25	G_2:2	0,001658	0,998342
26	G_2:2	0,001198	0,998802
27	G_1:1	0,999990	0,000010
28	---	0,906484	0,093516
29	G_1:1	0,999998	0,000002
30	G_1:1	0,999777	0,000223
31	G_2:2	0,000239	0,999761
32	G_1:1	0,998203	0,001797
33	G_1:1	0,996191	0,003809
34	G_2:2	0,021416	0,978584
35	G_1:1	1,000000	0,000000

Рисунок 18 – Апостериорные вероятности классификации

На основании таблиц, представленных на рисунках 17, 18, районы, не вошедшие в обучающие выборки, можно классифицировать следующим образом: районы Бугурусланский, Грачевский, Илекский, Курманаевский, Северный относятся к первому классу, т.е. социально-экономическое положение в этих районах хуже, чем в районах Абдулинский, Асекеевский, Октябрьский, Оренбургский, Первомайский, которые относятся ко второму классу.

2.4 Порядок выполнения лабораторной работы в пакете Stata

Для ввода исходных данных в пакет Stata необходимо выбрать пункты меню **Data**, **Data Editor** (рисунок 19). На экране появится пустая таблица, в которую

данные можно вводить вручную или скопировать через буфер обмена. При вводе данные следует учесть, что по умолчанию в пакете Stata дробная часть отделяется от целой части точкой, а не запятой.

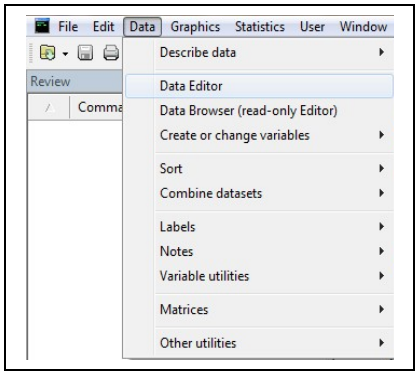
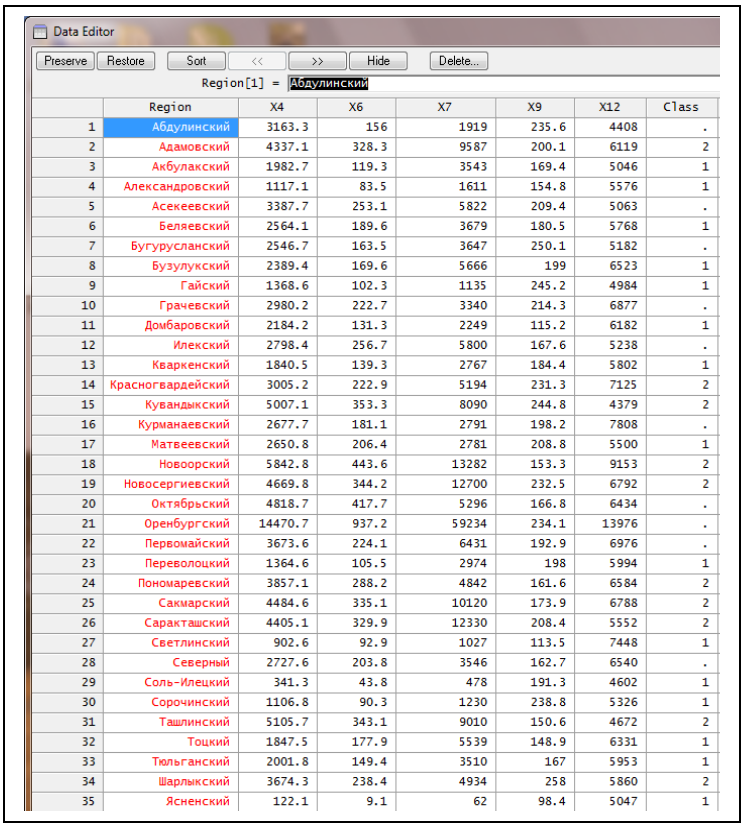


Рисунок 19 – Выбор пунктов меню для открытия таблицы с исходными данными

Таблица с исходными данными представлена на рисунке 20. Для удобства столбцам таблицы даны свои названия, это можно сделать двойным щелчком левой кнопки мыши на любой ячейке нужного столбца. Названия вводятся латинскими буквами.



	Region	X4	X6	X7	X9	X12	Class
1	Абдулинский	3163.3	156	1919	235.6	4408	.
2	Адамовский	4337.1	328.3	9587	200.1	6119	2
3	Акбулакский	1982.7	119.3	3543	169.4	5046	1
4	Александровский	1117.1	83.5	1611	154.8	5576	1
5	Асекеевский	3387.7	253.1	5822	209.4	5063	.
6	Беляевский	2564.1	189.6	3679	180.5	5768	1
7	Бугурусланский	2546.7	163.5	3647	250.1	5182	.
8	Бузулукский	2389.4	169.6	5666	199	6523	1
9	Гайский	1368.6	102.3	1135	245.2	4984	1
10	Грачевский	2980.2	222.7	3340	214.3	6877	.
11	Домбаровский	2184.2	131.3	2249	115.2	6182	1
12	Илекский	2798.4	256.7	5800	167.6	5238	.
13	Кваркенский	1840.5	139.3	2767	184.4	5802	1
14	Красногвардейский	3005.2	222.9	5194	231.3	7125	2
15	Кувандыкский	5007.1	353.3	8090	244.8	4379	2
16	Курманаевский	2677.7	181.1	2791	198.2	7808	.
17	Матвеевский	2650.8	206.4	2781	208.8	5500	1
18	Новоорский	5842.8	443.6	13282	153.3	9153	2
19	Новосергиевский	4669.8	344.2	12700	232.5	6792	2
20	Октябрьский	4818.7	417.7	5296	166.8	6434	.
21	Оренбургский	14470.7	937.2	59234	234.1	13976	.
22	Первомайский	3673.6	224.1	6431	192.9	6976	.
23	Переволоцкий	1364.6	105.5	2974	198	5994	1
24	Пономаревский	3857.1	288.2	4842	161.6	6584	2
25	Сакмарский	4484.6	335.1	10120	173.9	6788	2
26	Саракташский	4405.1	329.9	12330	208.4	5552	2
27	Светлинский	902.6	92.9	1027	113.5	7448	1
28	Северный	2727.6	203.8	3546	162.7	6540	.
29	Соль-Илецкий	341.3	43.8	478	191.3	4602	1
30	Сорочинский	1106.8	90.3	1230	238.8	5326	1
31	Ташлинский	5105.7	343.1	9010	150.6	4672	2
32	Тоцкий	1847.5	177.9	5539	148.9	6331	1
33	Тюльганский	2001.8	149.4	3510	167	5953	1
34	Шарлыкский	3674.3	238.4	4934	258	5860	2
35	Ясненский	122.1	9.1	62	98.4	5047	1

Рисунок 20 – Таблица с исходными данными в пакете Stata

Для реализации дискриминантного анализа в случае нормально распределенных классов с равными ковариационными матрицами необходимо выбрать пункты меню **Statistics**, **Multivariate analysis**, **Discriminant analysis**, **Linear (LDA)** (рисунок 21).

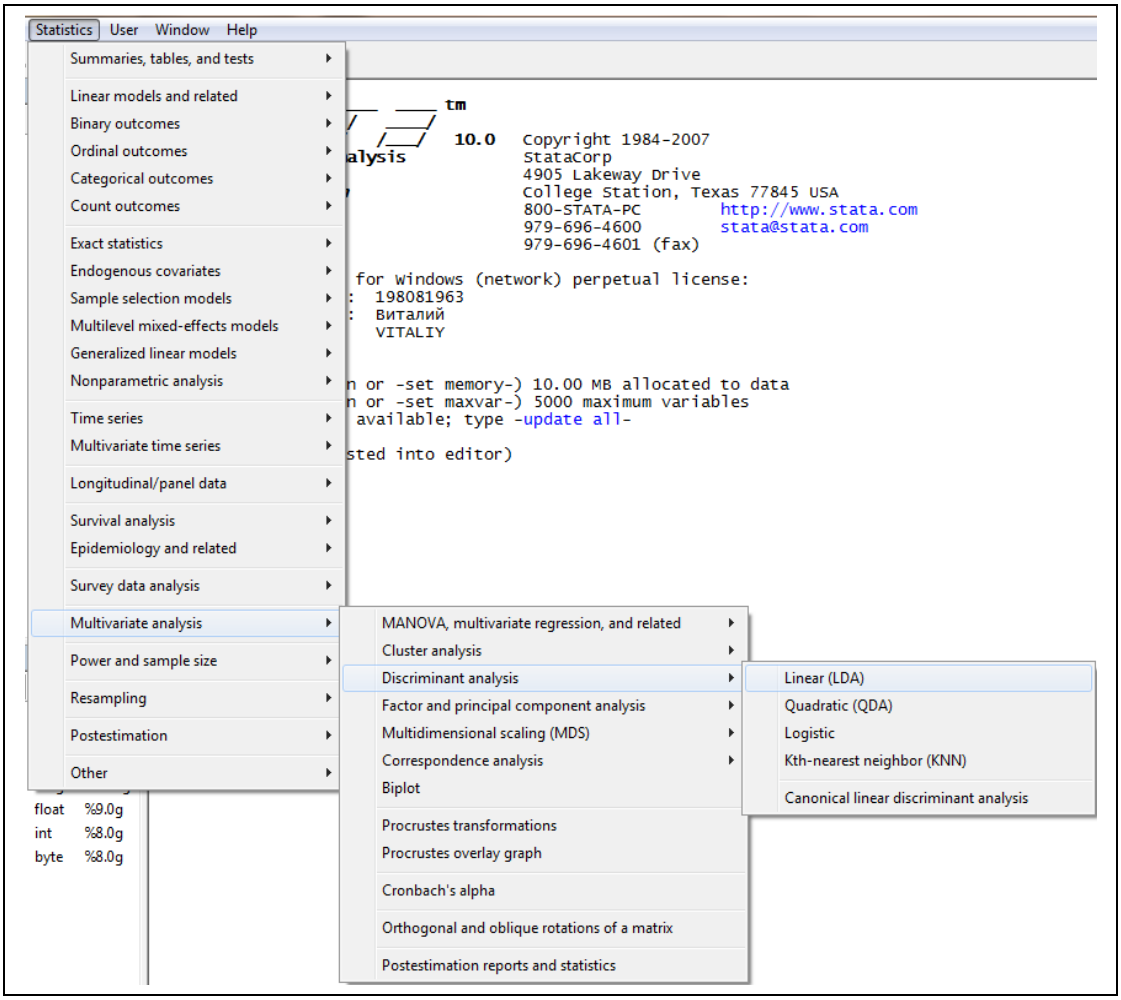


Рисунок 21 – Выбор пунктов меню для реализации дискриминантного анализа

На экране появится форма, представленная на рисунке 22. В поле **Variables** необходимо указать названия дискриминантных переменных (X4, X6, X7, X9, X12), а в поле **Group variable** – название столбца, в котором содержится номер класса (Class). В группе радио-кнопок **Group prior probabilities** устанавливается способ оценки удельных весов классов, выберем второй вариант – пропорционально объемам обучающих выборок.

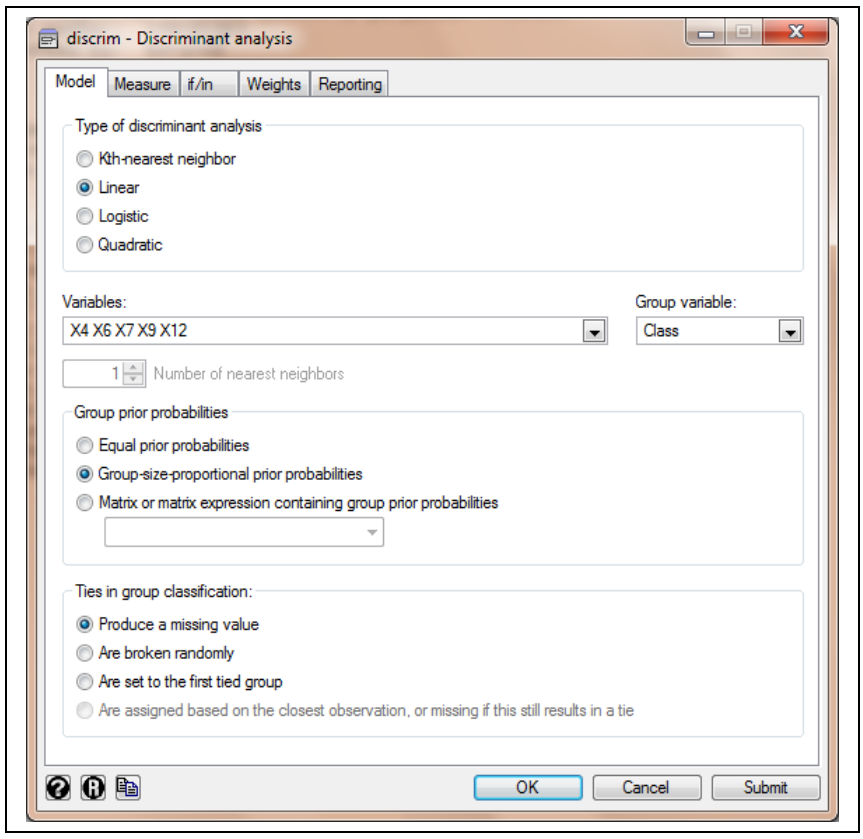


Рисунок 22 – Форма установки параметров дискриминантного анализа

После нажатия на кнопку ОК в окне результатов появится таблица с результатами классификации объектов из обучающих выборок и оценками априорных вероятностей классов. Вид экрана представлен на рисунке 23.

Linear discriminant analysis
Resubstitution classification summary

Key
Number Percent

True Class	Classified 1	2	Total
1	15 100.00	0 0.00	15 100.00
2	0 0.00	10 100.00	10 100.00
Total	15 60.00	10 40.00	25 100.00
Priors	0.6000	0.4000	

Рисунок 23 – Таблица с результатами классификации объектов из обучающих выборок

Анализируя таблицу, можно сделать вывод, что классификация объектов из обучающих выборок с помощью линейных дискриминантных функций Фишера полностью совпадает с исходной классификацией, а оценки априорных вероятностей составляют: $\hat{\pi}_1 = 0{,}6$, $\hat{\pi}_2 = 0{,}4$.

Для вывода на экран значений коэффициентов линейных дискриминантных функций Фишера и результатов классификации всех объектов необходимо выбрать пункты меню Statistics, Postestimation, Reports and statistics. Перечисленные пункты меню представлены на рисунке 24.

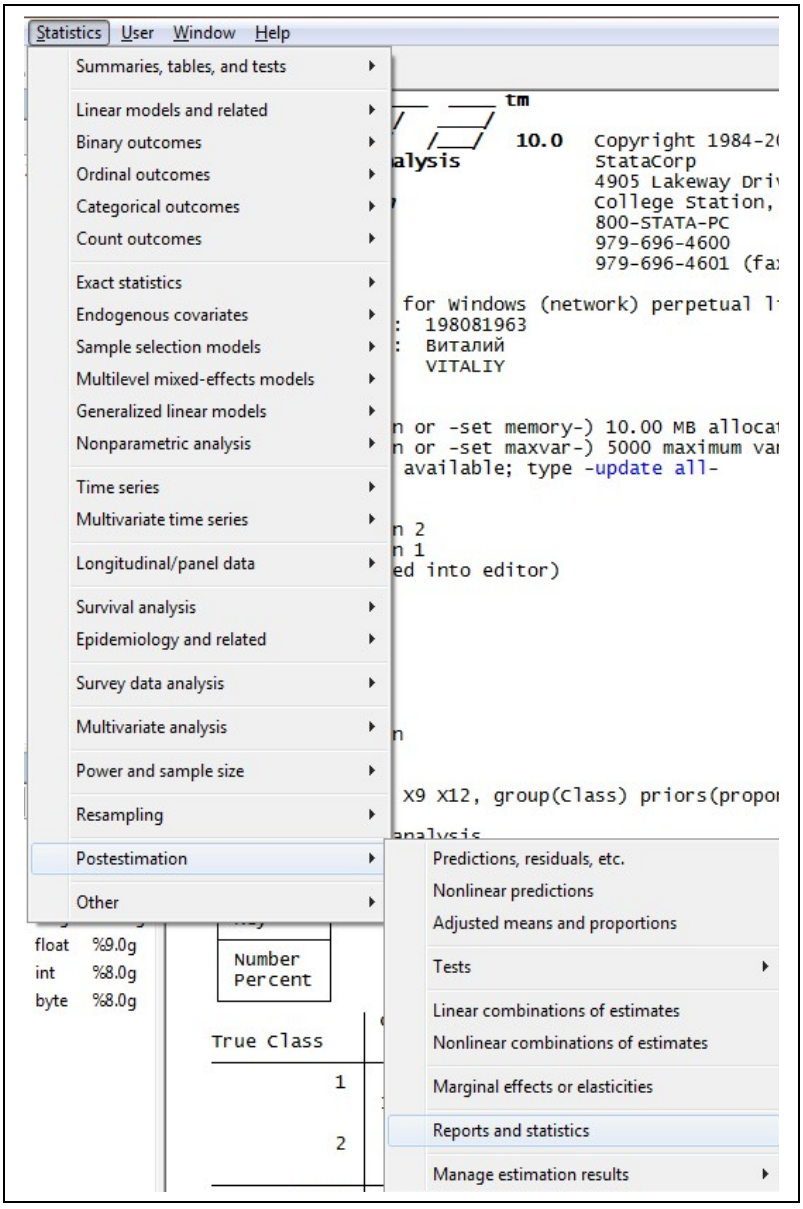


Рисунок 24 – Выбор пунктов меню для вывода результатов классификации

Для вывода коэффициентов линейных дискриминантных функций Фишера в форме, представленной на рисунке 25, необходимо из списка доступных отчетов выбрать **Classification (linear discriminant) functions (classfunctions)**. После нажатия на кнопке **ОК** на экране появятся результаты, представленные на рисунке 26.

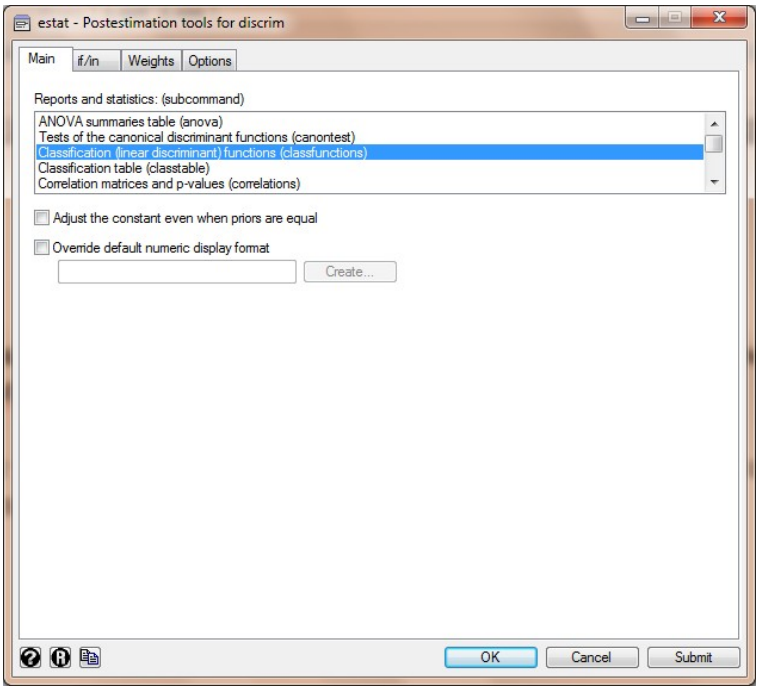


Рисунок 25 – Список доступных отчетов дискриминантного анализа

```
. estat classfunctions

Classification functions

                |     Class
                |         1          2
----------------+----------------------
             X4 |  .0139538   .0182476
             X6 |  -.165983  -.1623391
             X7 | -.0007167  -.0006558
             X9 |  .1671636   .1870107
            X12 |  .0093613   .0094696
          _cons | -42.06675  -60.94576
----------------+----------------------
         Priors |        .6          .4

.
```

Рисунок 26 – Коэффициенты линейных дискриминантных функций Фишера, рассчитанные в пакете Stata

Коэффициенты линейных дискриминантных функций Фишера, рассчитанные в пакете Stata, полностью совпадают с коэффициентами, рассчитанными в пакете Statistica и представленными на рисунке 15.

Для вывода на экран результатов классификации объектов с помощью линейных дискриминантных функций Фишера из списка доступных отчетов необходимо выбрать Classification listing (list). На экране появится форма с установками параметров вывода результатов, представленная на рисунке 27.

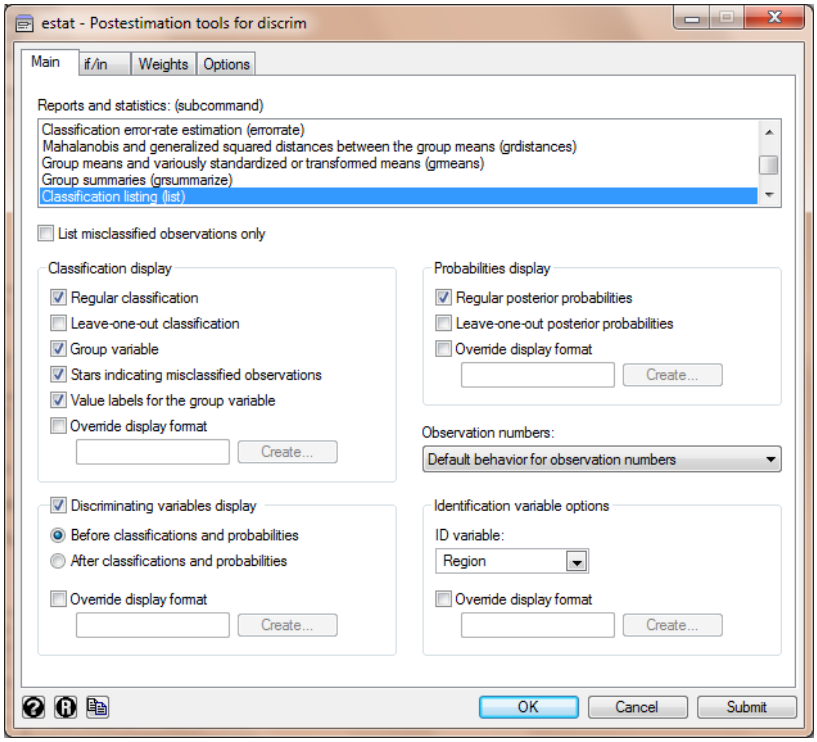


Рисунок 27 – Установка параметров вывода отчета с результатами классификации объектов

Для вывода в отчете значений дискриминантных переменных установим в форме галочку напротив **Discrimanation variables display**. В поле **ID variable** выберем название столбца, в котором содержится название района (**Region**). После нажатия на кнопку **ОК** в окне результатов появится отчет, представленный на рисунке 28.

```
. estat list, varlist id(Region)
```

	Data					Classification		Probabilities	
ID	X4	X6	X7	X9	X12	True	Class.	1	2
Абдулинский	3163.3	156	1919	235.6	4408	.	2	0.3677	0.6323
Адамовский	4337.1	328.3	9587	200.1	6119	2	2	0.0021	0.9979
Акбулакский	1982.7	119.3	3543	169.4	5046	1	1	0.9970	0.0030
Александровский	1117.1	83.5	1611	154.8	5576	1	1	1.0000	0.0000
Асекеевский	3387.7	253.1	5822	209.4	5063	.	2	0.1614	0.8386
Беляевский	2564.1	189.6	3679	180.5	5768	1	1	0.9398	0.0602
Бугурусланский	2546.7	163.5	3647	250.1	5182	.	1	0.8322	0.1678
Бузулукский	2389.4	169.6	5666	199	6523	1	1	0.9526	0.0474
Гайский	1368.6	102.3	1135	245.2	4984	1	1	0.9992	0.0008
Грачевский	2980.2	222.7	3340	214.3	6877	.	1	0.5175	0.4825
Домбаровский	2184.2	131.3	2249	115.2	6182	1	1	0.9973	0.0027
Илекский	2798.4	256.7	5800	167.6	5238	.	1	0.8431	0.1569
Кваркенский	1840.5	139.3	2767	184.4	5802	1	1	0.9976	0.0024
Красногвардейский	3005.2	222.9	5194	231.3	7125	2	2	0.3739	0.6261
Кувандыкский	5007.1	353.3	8090	244.8	4379	2	2	0.0001	0.9999
Курманаевский	2677.7	181.1	2791	198.2	7808	.	1	0.8548	0.1452
Матвеевский	2650.8	206.4	2781	208.8	5500	1	1	0.8625	0.1375
Новоорский	5842.8	443.6	13282	153.3	9153	2	2	0.0000	1.0000
Новосергиевский	4669.8	344.2	12700	232.5	6792	2	2	0.0002	0.9998
Октябрьский	4818.7	417.7	5296	166.8	6434	.	2	0.0005	0.9995
Оренбургский	14470.7	937.2	59234	234.1	13976	.	2	0.0000	1.0000
Первомайский	3673.6	224.1	6431	192.9	6976	.	2	0.0638	0.9362
Переволоцкий	1364.6	105.5	2974	198	5994	1	1	0.9996	0.0004
Пономаревский	3857.1	288.2	4842	161.6	6584	2	2	0.0498	0.9502
Сакмарский	4484.6	335.1	10120	173.9	6788	2	2	0.0017	0.9983
Саракташский	4405.1	329.9	12330	208.4	5552	2	2	0.0012	0.9988
Светлинский	902.6	92.9	1027	113.5	7448	1	1	1.0000	0.0000
Северный	2727.6	203.8	3546	162.7	6540	.	1	0.9065	0.0935
Соль-Илецкий	341.3	43.8	478	191.3	4602	1	1	1.0000	0.0000
Сорочинский	1106.8	90.3	1230	238.8	5326	1	1	0.9998	0.0002
Ташлинский	5105.7	343.1	9010	150.6	4672	2	2	0.0002	0.9998
Тоцкий	1847.5	177.9	5539	148.9	6331	1	1	0.9982	0.0018
Тюльганский	2001.8	149.4	3510	167	5953	1	1	0.9962	0.0038
Шарлыкский	3674.3	238.4	4934	258	5860	2	2	0.0214	0.9786
Ясненский	122.1	9.1	62	98.4	5047	1	1	1.0000	0.0000

Рисунок 28 – Отчет с результатами классификации в пакете Stata

Для того чтобы вывести на экран результаты классификации только тех районов, которые не вошли в обучающие выборки, необходимо установить галочку напротив List misclassified observations only в форме, представленной на рисунке 27. В этом случае отчет будет иметь вид, приведенный на рисунке 29.

```
. estat list, misclassified varlist id(Region)
```

	Data					Classification		Probabilities	
ID	X4	X6	X7	X9	X12	True	Class.	1	2
Абдулинский	3163.3	156	1919	235.6	4408	.	2	0.3677	0.6323
Асекеевский	3387.7	253.1	5822	209.4	5063	.	2	0.1614	0.8386
Бугурусланский	2546.7	163.5	3647	250.1	5182	.	1	0.8322	0.1678
Грачевский	2980.2	222.7	3340	214.3	6877	.	1	0.5175	0.4825
Илекский	2798.4	256.7	5800	167.6	5238	.	1	0.8431	0.1569
Курманаевский	2677.7	181.1	2791	198.2	7808	.	1	0.8548	0.1452
Октябрьский	4818.7	417.7	5296	166.8	6434	.	2	0.0005	0.9995
Оренбургский	14470.7	937.2	59234	234.1	13976	.	2	0.0000	1.0000
Первомайский	3673.6	224.1	6431	192.9	6976	.	2	0.0638	0.9362
Северный	2727.6	203.8	3546	162.7	6540	.	1	0.9065	0.0935

Рисунок 29 – Результаты классификации объектов, не вошедших в обучающие выборки

Результаты классификации, полученные в пакетах Statistica и Stata, полностью совпадают: Бугурусланский, Грачевский, Илекский, Курманаевский и Северный районы относятся к первому классу; Абдулинский, Асекеевский, Октябрьский, Оренбургский и Первомайский районы относятся ко второму классу.

Весь описанный алгоритм реализации дискриминантного анализа в пакете Stata можно выполнить с помощью четырех команд, которые в диалоговом режиме работы автоматически отражаются в специальном окне Review, представленном на рисунке 30.

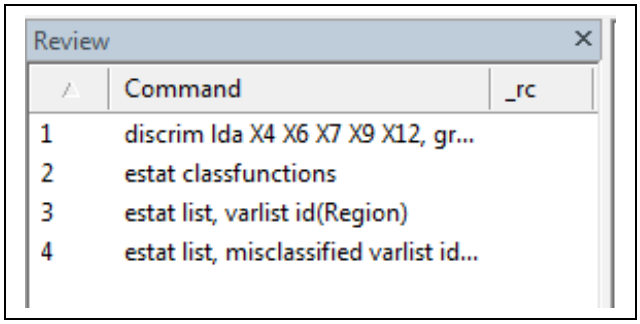
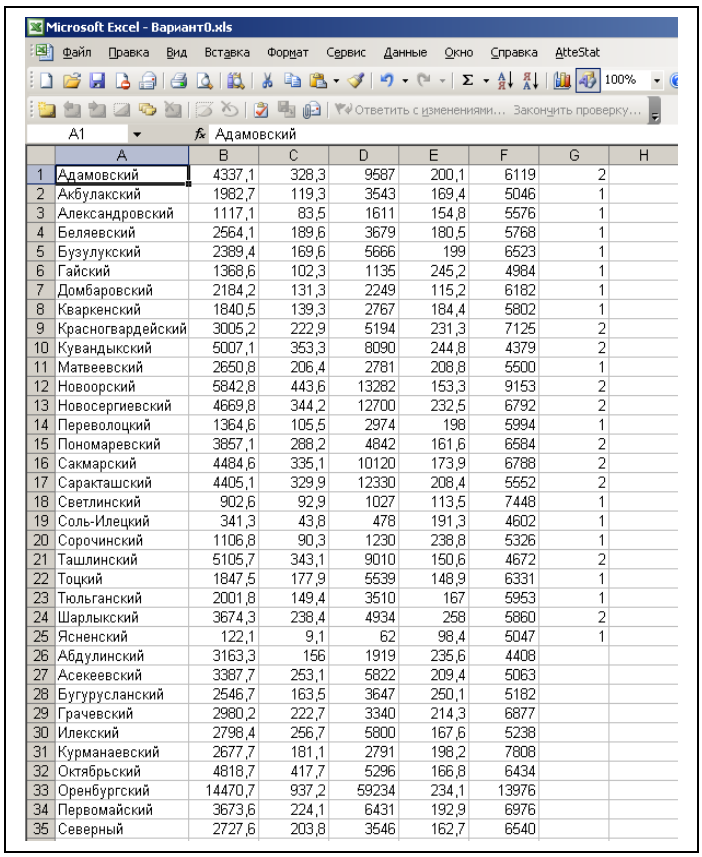


Рисунок 30 – Команды для реализации дискриминантного анализа

2.5 Порядок выполнения лабораторной работы с помощью надстройки AtteStat табличного процессора Microsoft Excel

Вид таблицы с исходными данными нулевого варианта в пакете Excel представлен на рисунке 31. В первых 25 строках введена статистическая информация по районам обучающих выборок, в следующих 10 строках – по районам, подлежащим классификации. В седьмом столбце (столбец G) указан номер класса (номер обучающей выборки).

Для реализации параметрического дискриминантного анализа с помощью надстройки AtteStat необходимо выбрать пункт основного меню **AtteStat**, подпункты **Модуль PRT – Распознавание образов**, **Распознавание образов**. Вид экрана представлен на рисунке 32.



Microsoft Excel - Вариант0.xls

Файл Правка Вид Вставка Формат Сервис Данные Окно Справка AtteStat

A1 Адамовский

	A	B	C	D	E	F	G	H
1	Адамовский	4337,1	328,3	9587	200,1	6119	2	
2	Акбулакский	1982,7	119,3	3543	169,4	5046	1	
3	Александровский	1117,1	83,5	1611	154,8	5576	1	
4	Беляевский	2564,1	189,6	3679	180,5	5768	1	
5	Бузулукский	2389,4	169,6	5666	199	6523	1	
6	Гайский	1368,6	102,3	1135	245,2	4984	1	
7	Домбаровский	2184,2	131,3	2249	115,2	6182	1	
8	Кваркенский	1840,5	139,3	2767	184,4	5802	1	
9	Красногвардейский	3005,2	222,9	5194	231,3	7125	2	
10	Кувандыкский	5007,1	353,3	8090	244,8	4379	2	
11	Матвеевский	2650,8	206,4	2781	208,8	5500	1	
12	Новоорский	5842,8	443,6	13282	153,3	9153	2	
13	Новосергиевский	4669,8	344,2	12700	232,5	6792	2	
14	Переволоцкий	1364,6	105,5	2974	198	5994	1	
15	Пономаревский	3857,1	288,2	4842	161,6	6584	2	
16	Сакмарский	4484,6	335,1	10120	173,9	6788	2	
17	Саракташский	4405,1	329,9	12330	208,4	5552	2	
18	Светлинский	902,6	92,9	1027	113,5	7448	1	
19	Соль-Илецкий	341,3	43,8	478	191,3	4602	1	
20	Сорочинский	1106,8	90,3	1230	238,8	5326	1	
21	Ташлинский	5105,7	343,1	9010	150,6	4672	2	
22	Тоцкий	1847,5	177,9	5539	148,9	6331	1	
23	Тюльганский	2001,8	149,4	3510	167	5953	1	
24	Шарлыкский	3674,3	238,4	4934	258	5860	2	
25	Ясненский	122,1	9,1	62	98,4	5047	1	
26	Абдулинский	3163,3	156	1919	235,6	4408		
27	Асекеевский	3387,7	253,1	5822	209,4	5063		
28	Бугурусланский	2546,7	163,5	3647	250,1	5182		
29	Грачевский	2980,2	222,7	3340	214,3	6877		
30	Илекский	2798,4	256,7	5800	167,6	5238		
31	Курманаевский	2677,7	181,1	2791	198,2	7808		
32	Октябрьский	4818,7	417,7	5296	166,8	6434		
33	Оренбургский	14470,7	937,2	59234	234,1	13976		
34	Первомайский	3673,6	224,1	6431	192,9	6976		
35	Северный	2727,6	203,8	3546	162,7	6540		

Рисунок 31 – Исходные данные в пакете Excel

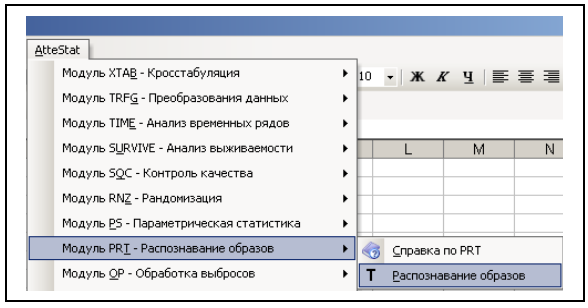


Рисунок 32 – Выбор пунктов меню в пакете Excel

Для реализации линейного дискриминантного анализа Фишера необходимо заполнить появившуюся на экране форму «Распознавание образов с обучением»: в поле «Интервал обучающей выборки» вводится диапазон статистических данных по районам, составляющим обучающие выборки; в поле «Интервал номеров классов или оценок» вводится диапазон ячеек, в которых введены номера классов; в поле «Интервал вывода результатов» указывается ячейка, с которой начнется вывод результатов. Вид заполненной формы представлен на рисунке 33.

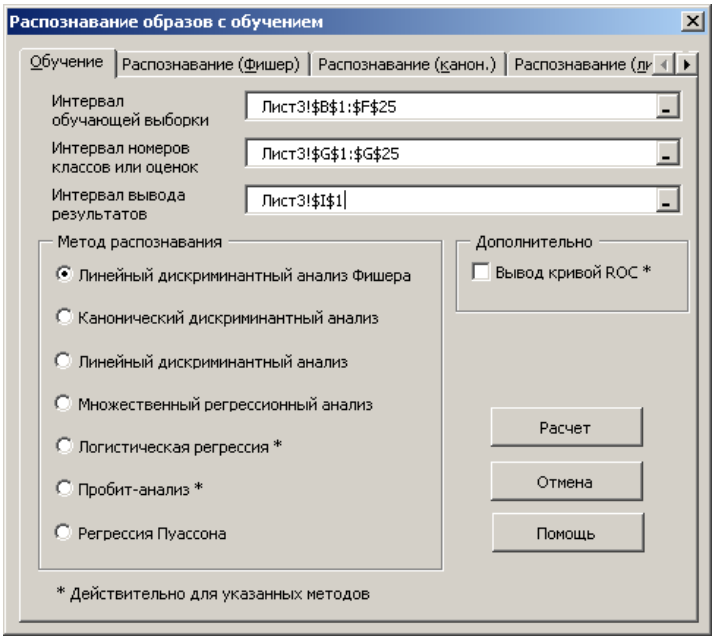


Рисунок 33 – Образец заполнения формы «Распознавание образов с обучением» для реализации линейного дискриминантного анализа Фишера (страница «Обучение»)

С помощью кнопки **Расчет** в таблице с исходными данными появятся результаты линейного дискриминантного анализа Фишера, представленные на рисунке 34.

I	J	K	L	M
Число объектов обучающей выборки				
25				
Число параметров				
5				
Число классов				
2				
Численности классов				
15				
10				
Линейный дискриминантный анализ по Фишеру				
Качество распознавания, %				
100				
Простые классифицирующие функции				
(в столбце - константа, коэффициенты)				
-41,5559	-60,0295			
0,013954	0,018248			
-0,16598	-0,16234			
-0,00072	-0,00066			
0,167164	0,187011			
0,009361	0,00947			

Рисунок 34 – Результаты реализации линейного дискриминантного анализа Фишера с помощью надстройки AtteStat пакета Excel

Таким образом, линейные дискриминантные функции Фишера имеют вид:

$$\varphi_1(x_4, x_6, x_7, x_9, x_{12}) = -41{,}5559 + 0{,}01395x_4 - 0{,}16598x_6 - \\ - 0{,}00072x_7 + 0{,}16716x_9 + 0{,}00936x_{12}; \quad (15)$$

$$\varphi_2(x_4, x_6, x_7, x_9, x_{12}) = -60{,}0295 + 0{,}01825x_4 - 0{,}16234x_6 - \\ - 0{,}00066x_7 + 0{,}18701x_9 + 0{,}00947x_{12}. \quad (16)$$

Для осуществления классификации районов на основе функций (15), (16) необходимо снова открыть форму «Распознавание образов с обучением» и выбрать страницу «Распознавание (Фишер)». В поле «Объект» вводится диапазон ячеек со статистическими данными, характеризующими район, которых подлежит классификации, например, Абдулинский; в поле «Простые классифицирующие функции» вводится диапазон ячеек с коэффициентами дискриминантных функций Фишера. Заполненная форма приведена на рисунке 35.

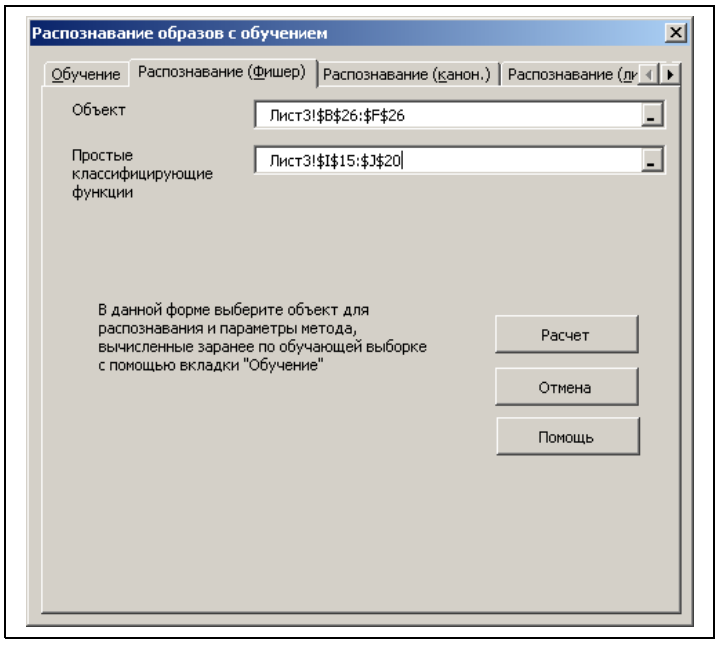


Рисунок 35 – Образец заполнения формы «Распознавание образов с обучением» (страница «Распознавание (Фишер)»)

После нажатия кнопки **Расчет** на экране появится форма с результатом классификации Абдулинского района, представленная на рисунке 36.

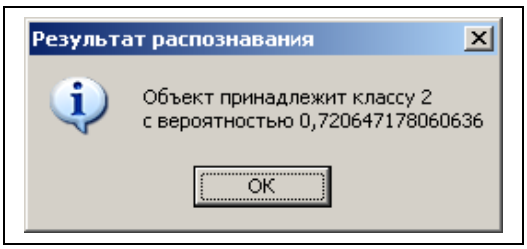


Рисунок 36 – Результат классификации Абдулинского района

Таким образом, Абдулинский район с вероятностью 0,72 следует отнести ко второму классу. Аналогичным образом осуществляется классификация остальных девяти районов Оренбургской области. Результаты удобно свести в таблицу 1.

Таблица 1 – Результаты классификации районов с помощью линейных дискриминантных функций Фишера с помощью надстройки AtteStat пакета Excel

Номер района	**Наименование района**	**Номер класса**	**Вероятность**
1	Абдулинский	2	0,72
5	Асекеевский	2	0,89
7	Бугурусланский	1	0,77
10	Грачевский	2	0,58
12	Илекский	1	0,78
16	Курманаевский	1	0,80
20	Октябрьский	2	0,99
21	Оренбургский	2	1,00
22	Первомайский	2	0,96
28	Северный	1	0,86

Отличие в результатах классификации Грачевского района можно объяснить некоторым различием в коэффициентах дискриминантных функций (12), (13) и (14), (15).

Реализованный линейный дискриминантный анализ Фишера основан на предположении нормального закона распределения классов с равными ковариационными матрицами. При необоснованном объявлении ковариационных матриц статистически неразличимыми в результате реализации линейного дискриминантного анализа Фишера могут оказаться отброшенными важные индивидуальные черты, имеющие большое значение для хорошей дискриминации.

В этом случае в надстройке AtteStat пакета Excel реализован линейный дискриминантный анализ, в основе которого лежит правило классификации (4) применительно к нормально распределенным классам. При этом необходимо проверить, чтобы число объектов в каждой обучающей выборки было хотя бы на 2 единицы больше чем число признаков.

Для реализации линейного дискриминантного анализа необходимо в форме «Распознавание образов с обучением» выбрать «Линейный дискриминантный анализ». Поля «Интервал обучающей выборки», «Интервал номеров классов или оценок», «Интервал вывода результатов» заполняются аналогично тому, как описано выше. Вид заполненной формы представлен на рисунке 37.

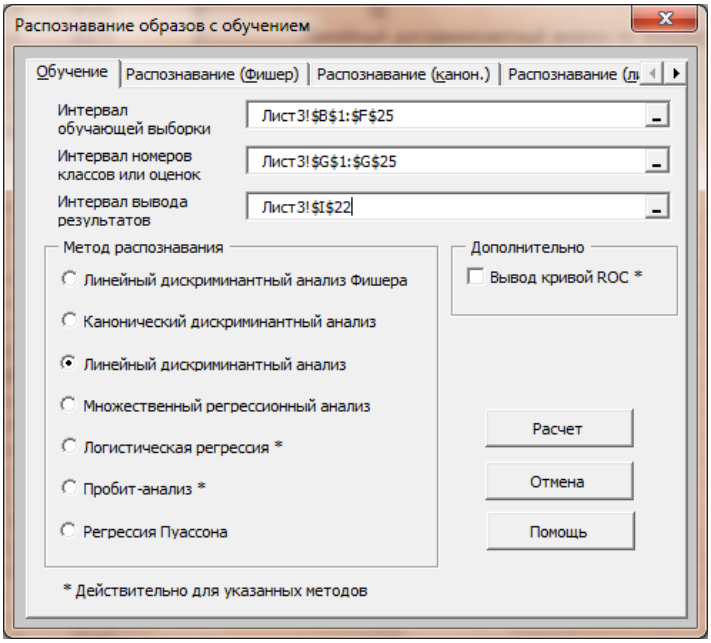
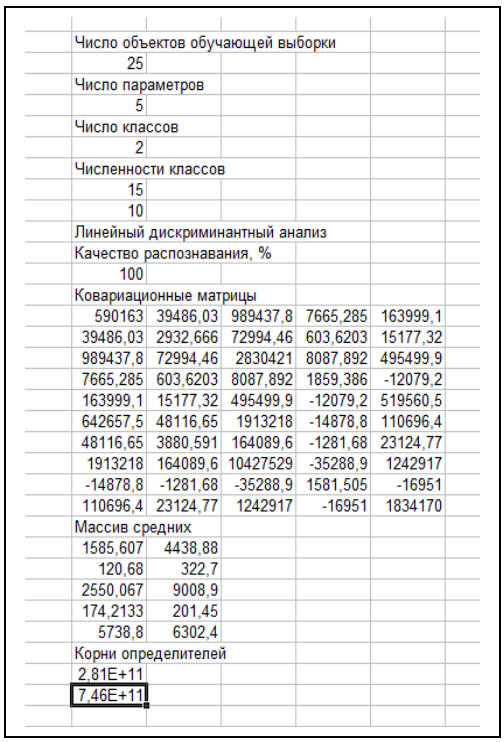


Рисунок 37 – Образец заполнения формы «Распознавание образов с обучением» для реализации линейного дискриминантного анализа (страница «Обучение»)

С помощью кнопки **Расчет** в таблице с исходными данными появятся результаты линейного дискриминантного анализа, представленные на рисунке 38.



Число объектов обучающей выборки					
25					
Число параметров					
5					
Число классов					
2					
Численности классов					
15					
10					
Линейный дискриминантный анализ					
Качество распознавания, %					
100					
Ковариационные матрицы					
590163	39486,03	989437,8	7665,285	163999,1	
39486,03	2932,666	72994,46	603,6203	15177,32	
989437,8	72994,46	2830421	8087,892	495499,9	
7665,285	603,6203	8087,892	1859,386	-12079,2	
163999,1	15177,32	495499,9	-12079,2	519560,5	
642657,5	48116,65	1913218	-14878,8	110696,4	
48116,65	3880,591	164089,6	-1281,68	23124,77	
1913218	164089,6	10427529	-35288,9	1242917	
-14878,8	-1281,68	-35288,9	1581,505	-16951	
110696,4	23124,77	1242917	-16951	1834170	
Массив средних					
1585,607	4438,88				
120,68	322,7				
2550,067	9008,9				
174,2133	201,45				
5738,8	6302,4				
Корни определителей					
2,81E+11					
7,46E+11					

Рисунок 38 – Результаты реализации линейного дискриминантного анализа с помощью надстройки AtteStat пакета Excel

С помощью надстройки AtteStat на основе обучающих выборок рассчитаны оценки параметров нормально распределенных классов - оценки ковариационных матриц и векторов математических ожиданий, а также корни квадратные из определителей оценок ковариационных матриц. Эти результаты будут использованы программой для осуществления классификации районов на основе правила (4). Для этого необходимо снова открыть форму «Распознавание образов с обучением» и выбрать страницу «Распознавание (лин.)». В поле «Объект» вводится диапазон ячеек со статистическими данными, характеризующими район, которых подлежит классификации, например, Асекеевский; в поле «Ковариационные матрицы» вводится диапазон ячеек с элементами ковариационных матриц; в поле «Массив средних» вводится диапазон ячеек, содержащих средние арифметические значения признаков в классах; в поле «Корни определителей» вводится диапазон ячеек со значениями корней квадратных из определителей оценок ковариационных матриц. Заполненная форма приведена на рисунке 39.

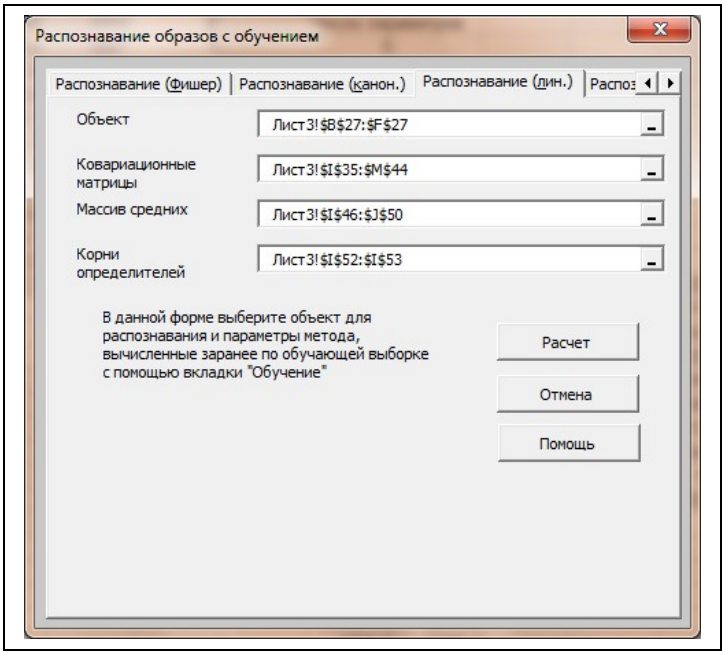


Рисунок 39 – Образец заполнения формы «Распознавание образов с обучением» (страница «Распознавание (лин.)»)

После нажатия кнопки **Расчет** на экране появится форма с результатом классификации Асекеевского района, представленная на рисунке 40.

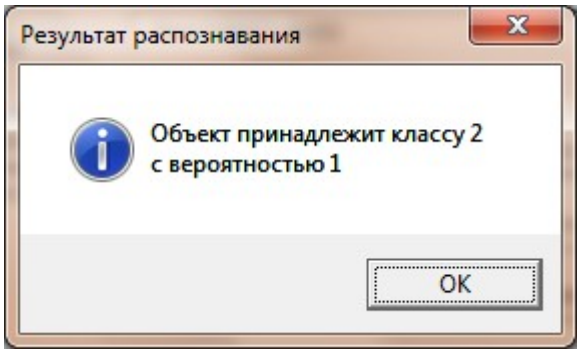


Рисунок 40 – Результат классификации Асекеевского района

Таким образом, Асекеевский район с вероятностью 1 следует отнести ко второму классу. Аналогичным образом осуществляется классификация остальных девяти районов Оренбургской области. Результаты классификации сведены в таблицу 2.

Таблица 2 – Результаты классификации районов с помощью линейных дискриминантных функций с помощью надстройки AtteStat пакета Excel

Номер класса	**Районы**
Первый класс	Грачевский, Илекский, Октябрьский, Первомайский, Северный
Второй класс	Абдулинский, Асекеевский, Бугурусланский, Курманаевский, Оренбургский

В результатах классификации, полученных с помощью двух методов, имеются заметные различия. В первом методе (линейный дискриминантный анализ Фишера) предполагается равенство ковариационных матриц двух классов, во втором методе такого предположения не делается. Полученные различия в результатах классификации свидетельствуют о необоснованности предположения равенства ковариационных матриц двух классов, на что указывают также оценки ковариационных матриц, рассчитанные с помощью надстройки AtteStat пакета Excel (смотри рисунок 38):

$$\hat{\Sigma}_1 = \begin{pmatrix} 590163 & 39486 & 989437 & 7665 & 163999 \\ 39486 & 2933 & 72994 & 604 & 15177 \\ 989437 & 72994 & 2830421 & 8088 & 495500 \\ 7665 & 604 & 8088 & 1859 & -12079 \\ 163999 & 15177 & 495500 & -12079 & 519560 \end{pmatrix};$$

$$\hat{\Sigma}_2 = \begin{pmatrix} 642657 & 48117 & 1913218 & -14879 & 110696 \\ 48117 & 3881 & 164090 & -1282 & 23125 \\ 1913218 & 164090 & 10427529 & -35289 & 1242917 \\ -14879 & -1282 & -35289 & 1581 & -16951 \\ 110696 & 23125 & 1242917 & -16951 & 1834170 \end{pmatrix}.$$

2.6 Содержание письменного отчета

Отчет должен быть оформлен на листах формата А4 с титульным листом, оформленным соответствующим образом, и содержать следующее:

1) постановку задачи;
2) исходные данные для анализа;
3) краткое изложение теории;
4) результаты выполнения лабораторной работы.

2.7 Вопросы к защите лабораторной работы

1) Сформулируйте постановку задачи лабораторной работы
2) Каким методом классификации решалась задача и чем обусловлен этот выбор?
3) Сформулируйте, в чем суть выбранного метода решения задачи
4) Какое программное обеспечение использовалось для решения задачи и какие правила классификации реализованы в каждом из инструментальных средств?
5) Объясните, зачем нужны в лабораторной работе обучающие выборки. Можно ли при реализации дискриминантного анализа обойтись без них?
6) Какому условию должно удовлетворять количество объектов в обучающей выборке для реализации линейного дискриминантного анализа?
7) На основе какой информации можно дать характеристику классам?
8) Объясните, каким образом проводить классификацию объектов на основе результатов, выдаваемых пакетом Statistica в форме таблиц, представленных на рисунках 17, 18?
9) Запишите формулу для расчета квадрата расстояния Махаланобиса от объекта x_v до центра каждого из классов в лабораторной работе [5, с. 492]
10) Осуществите классификацию района $x_v = (1000; 1000; 1000; 1000; 1000)^T$
11) В каком случае качество распознавания объектов будет меньше 100%?
12) Продемонстрируйте, каким образом изменятся алгоритм работы с пакетами, выдаваемые результаты и их интерпретация в следующих случаях:

- уменьшилось количество объектов в первой обучающей выборке на один район;
- количество признаков сократилось до первых трех;
- увеличилось количество обучающих выборок на одну.

13) Есть ли различия в результатах классификации районов при различных допущениях о характере распределения классов? На что это указывает?

Список использованных источников

1 Боровиков, В.П. STATISTICA – Статистический анализ и обработка данных в среде Windows / В.П. Боровиков, И.П. Боровиков. – М.: Инф. изд. дом «Филин», 1998. – 608 с.

2 Дубров, А.М. Многомерные статистические методы: учебник / А.М. Дубров, В.С. Мхитарян, Л.И. Трошин. – М.: Финансы и статистика, 1998. – 352 с.

3 Дуброва, Т.А. Дискриминантный анализ в системе «STATISTICA»: учебное пособие / Т.А. Дуброва, А.Г. Бажин, Л.П. Бакуменко. – М.: Московский государственный университет экономики, статистики и информатики, 2000. – 57 с.

4 Областной статистический ежегодник: стат.сб. / Территориальный орган Федеральной службы государственной статистики по Оренбургской области. – Оренбург, 2009. – 500 с.

5 Айвазян, С.А. Прикладная статистика. Основы эконометрики: учебник для вузов: в 2 т. / С.А. Айвазян, В.С. Мхитарян. – М.: ЮНИТИ-ДАНА, 2001. – Т. 1: Теория вероятностей и прикладная статистика. – 656 с.

6 Реннер, А.Г. Параметрический дискриминантный анализ: методические указания к лабораторному практикуму и самостоятельной работе студентов / А.Г. Реннер, О.С. Чудинова.– Оренбург: ГОУ ОГУ, 2010. – 19 с.

7 Сошникова, Л.А. Многомерный статистический анализ в экономике: учеб. пособие для вузов / Л.А. Сошникова, В.Н. Тамашевич, Г.Е. Уебе, М. Шефер. – М.: ЮНИТИ, 1999. – 598 с.

8 Тюрин, Ю.Н. Статистический анализ данных на компьютере / Ю.Н. Тюрин, А.А. Макаров; под ред. В.Э. Фигурнова. – М.: ИНФРА-М, 1998. – 528 с.

Приложение А

(обязательное)

Исходные данные для анализа

Таблица А.1 – Обозначение и наименование показателей

Обозначение	Наименование показателя
x_1	инвестиции в основной капитал, тыс.руб.
x_2	инвестиции в основной капитал на душу населения, руб.
x_3	инвестиции, направленные в жилищное хозяйство, тыс.руб.
x_4	инвестиции, направленные в жилищное хозяйство, на душу населения, руб.
x_5	ввод в действие жилых домов, кв.м
x_6	ввод в действие жилых домов на 1000 человек населения, кв.м
x_7	ввод в действие жилых домов, построенных населением за свой счет и с помощью кредитов, кв.м
x_8	площадь жилищ, приходящаяся в среднем на одного жителя, кв.м
x_9	обеспеченность населения собственными легковыми автомобилями в расчете на 1000 населения, штук
x_{10}	фонд оплаты труда работников, млн. руб.
x_{11}	среднемесячная начисленная заработная плата работников, руб.
x_{12}	удельный вес убыточных организаций, в % от общего числа организаций
x_{13}	число предприятий и организаций строительства
x_{14}	оборот розничной торговли на душу населения, руб.

Таблица А.2 – Значения социально-экономических показателей, характеризующих районы Оренбургской области, за 2007 год

Номер района	Наименование района	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
1	Абдулинский	109189	8877	38908	3163,3	1919	156,0	1919	24,0	235,6	119,6	4408	18,2	1	8378
2	Адамовский	646948	21565	130112	4337,1	9848	328,3	9587	19,6	200,1	638,5	6119	5,6	7	11051
3	Акбулакский	229355	7722	58886	1982,7	3543	119,3	3543	17,5	169,4	334,6	5046	26,7	14	7194
4	Александровский	73185	3792	21560	1117,1	1611	83,5	1611	18,8	154,8	264,1	5576	26,7	4	11919
5	Асекеевский	259212	11270	77916	3387,7	5822	253,1	5822	19,6	209,4	301,1	5063	29,4	4	8043
6	Беляевский	129238	6662	49744	2564,1	3679	189,6	3679	19,2	180,5	280,2	5768	37,5	5	8629
7	Бугурусланский	288477	12936	56792	2546,7	3647	163,5	3647	20,4	250,1	271,5	5182	50,0	12	13728
8	Бузулукский	176376	5281	79807	2389,4	5666	169,6	5666	20,0	199,0	545,5	6523	36,8	10	6121
9	Гайский	124545	11220	15191	1368,6	1135	102,3	1135	20,8	245,2	128,4	4984	28,6	1	12275
10	Грачевский	178909	11927	44703	2980,2	3340	222,7	3340	22,7	214,3	297,6	6877	18,2	3	9073
11	Домбаровский	113887	6090	40844	2184,2	2456	131,3	2249	20,4	115,2	258,1	6182	37,5	10	8897
12	Илекский	201464	7069	79755	2798,4	7316	256,7	5800	18,8	167,6	353,2	5238	25,3	7	8187
13	Кваркенский	164722	7661	39571	1840,5	2995	139,3	2767	19,4	184,4	363,9	5802	42,1	3	8495
14	Красногвардейский	205538	8821	70022	3005,2	5194	222,9	5194	21,0	231,3	394,7	7125	16,7	8	9672
15	Кувандыкский	200475	8754	114663	5007,1	8090	353,3	8090	18,2	244,8	227,4	4379	18,8	3	4494
16	Курманаевский	119348	5938	53821	2677,7	3641	181,1	2791	21,7	198,2	398,3	7808	40,0	8	8203
17	Матвеевский	80125	5451	38967	2650,8	3034	206,4	2781	20,9	208,8	209,8	5500	33,3	5	13456
18	Новоорский	1113783	34806	186969	5842,8	14196	443,6	13282	21,3	153,3	749,0	9153	25,0	24	14059
19	Новосергиевский	451312	12231	172317	4669,8	12700	344,2	12700	21,3	232,5	806,7	6792	27,8	12	22151
20	Октябрьский	357604	16108	106976	4818,7	9272	417,7	5296	21,9	166,8	439,7	6434	17,6	9	12860
21	Оренбургский	4542065	62220	1056362	14470,7	68415	937,2	59234	20,5	234,1	4389,9	13976	20,4	290	25230
22	Первомайский	251218	8753	105431	3673,6	6431	224,1	6431	18,2	192,9	534,4	6976	18,8	34	11456
23	Переволоцкий	174074	5881	40391	1364,6	3122	105,5	2974	19,8	198,0	419,0	5994	9,1	11	11327
24	Пономаревский	95263	5670	64800	3857,1	4842	288,2	4842	22,9	161,6	223,3	6584	83,3	10	16004
25	Сакмарский	432616	14325	135436	4484,6	10120	335,1	10120	17,9	173,9	463,0	6788	52,4	24	10789

Продолжение таблицы А.2

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
26	Саракташский	482734	11200	189860	4405,1	14219	329,9	12330	18,3	208,4	654,5	5552	10,7	20	12486
27	Светлинский	91578	5324	15524	902,6	1598	92,9	1027	19,9	113,5	425,8	7448	7,7	5	9309
28	Северный	148732	8548	47460	2727,6	3546	203,8	3546	21,2	162,7	306,1	6540	36,4	5	16139
29	Соль-Илецкий	93681	3419	9351	341,3	1199	43,8	478	16,6	191,3	274,7	4602	41,7	2	16105
30	Сорочинский	113378	7268	17266	1106,8	1409	90,3	1230	21,4	238,8	243,1	5326	13,3	10	5818
31	Ташлинский	436789	16359	136321	5105,7	9162	343,1	9010	21,0	150,6	439,0	4672	17,6	3	13005
32	Тоцкий	170168	4212	74639	1847,5	7188	177,9	5539	18,1	148,9	398,3	6331	40,0	12	12484
33	Тюльганский	150351	6398	47043	2001,8	3510	149,4	3510	19,8	167,0	360,8	5953	40,0	5	11844
34	Шарлыкский	238024	11499	76059	3674,3	4934	238,4	4934	21,0	258,0	331,7	5860	9,1	11	16074
35	Ясненский	43901	6456	830	122,1	62	9,1	62	17,5	98,4	115,8	5047	25,7	10	6326

Таблица А.3 – Варианты заданий

Номер варианта	Набор показателей	Количество обучающих выборок	Обучающие выборки	
			Номер обучающей выборки	Порядковые номера районов
(1)	(2)	(3)	(4)	(5)
0	x_4, x_6, x_7, x_9, x_{12}	2	1	3, 4, 6, 8, 9, 11, 13, 17, 23, 27, 29, 30, 32, 33, 35
			2	2, 14, 15, 18, 19, 24, 25, 26, 31, 34
1	x_4, x_6, x_7, x_9, x_{12}	2	1	2, 5, 15, 18, 20, 22, 25, 26, 31, 34
			2	1, 4, 6, 7, 9, 10, 13, 17, 23, 27, 28, 30, 32, 33, 35
2	x_4, x_6, x_7, x_9, x_{12}	2	1	2, 5, 14, 15, 18, 19, 20, 24, 34
			2	1, 3, 7, 8, 9, 10, 11, 16, 23, 27, 28, 29, 33, 35
3	x_3, x_4, x_5, x_7, x_{15}	3	1	1, 6, 8, 10, 12, 14, 16, 17, 24, 28, 33, 34
			2	4, 13, 23, 27, 29, 30, 35
			3	2, 15, 18, 19, 20, 26, 31
4	x_3, x_4, x_5, x_7, x_{15}	3	1	1, 3, 7, 9, 10, 11, 16, 17, 24, 32, 33, 34
			2	9, 13, 23, 27, 29, 30, 35
			3	18, 19, 20, 22, 25, 26, 31
5	x_3, x_4, x_5, x_7, x_{15}	3	1	3, 5, 6, 7, 8, 14, 16, 17, 24, 28
			2	4, 9, 13, 23, 27, 30, 35
			3	2, 15, 18, 19, 20, 22, 25, 26
6	x_3, x_4, x_5, x_7, x_{15}	2	1	1, 3, 5, 6, 8, 10, 12, 14, 16, 17, 23, 24, 27, 29, 30, 32, 33, 34
			2	2, 18, 19, 22, 25, 26, 31
7	x_3, x_4, x_5, x_7, x_{15}	2	1	2, 15, 18, 20, 22, 25, 31
			2	1, 3, 4, 5, 6, 8, 9, 11, 14, 17, 23, 24, 27, 28, 32, 33
8	x_1, x_5, x_7, x_{12}, x_{13}	2	1	2, 18, 19, 20, 25, 26, 31
			2	1, 3, 5, 9, 10, 13, 14, 15, 16, 17, 22, 24, 27, 28, 30, 32, 33, 35
9	x_1, x_5, x_7, x_{12}, x_{13}	2	1	4, 5, 6, 7, 11, 12, 13, 14, 15, 16, 22, 23, 27, 28, 29, 30, 34, 35
			2	2, 18, 19, 20, 25, 26, 31
10	x_1, x_3, x_5, x_{11}, x_{12}	2	1	1, 4, 6, 8, 10, 12, 14, 16, 17, 23, 24, 27, 28, 29, 30, 32, 33, 35
			2	2, 18, 20, 22, 25, 26, 31

Продолжение таблицы А.3

(1)	(2)	(3)	(4)	(5)
11	x_1, x_3, x_5, x_{11}, x_{12}	2	1	1, 3, 5, 6, 8, 10, 12, 14, 16, 23, 24, 27, 28, 29, 30, 32, 33, 34,
			2	2, 18, 19, 20, 22, 26, 31
12	x_2, x_4, x_6, x_9, x_{12}	2	1	1, 3, 6, 7, 9, 11, 13, 16, 17, 23, 27, 29, 32, 33, 35
			2	2, 5, 15, 19, 20, 22, 24, 25, 31, 34
13	x_2, x_4, x_6, x_9, x_{12}	2	1	4, 6, 8, 9, 13, 16, 23, 27, 28, 29, 30, 32, 33, 35
			2	5, 14, 15, 18, 19, 22, 24, 25, 26, 31, 34
14	x_2, x_4, x_6, x_9, x_{12}	2	1	14, 15, 18, 19, 20, 22, 24, 25, 26, 31
			2	1, 3, 4, 7, 9, 11, 12, 16, 17, 23, 27, 30, 32, 33, 35
15	x_4, x_6, x_8, x_{12}, x_{17}	2	1	1, 4, 6, 8, 11, 13, 17, 23, 27, 29, 30, 32, 33, 35
			2	2, 10, 14, 16, 18, 20, 22, 24, 25, 26, 28
16	x_4, x_6, x_8, x_{12}, x_{17}	2	1	3, 5, 7, 8, 9, 11, 12, 13, 17, 23, 27, 29, 30, 32, 33, 35
			2	15, 16, 18, 19, 20, 22, 24, 25, 28, 31, 34
17	x_4, x_6, x_8, x_{12}, x_{17}	2	1	1, 3, 4, 5, 6, 7, 8, 9, 11, 12, 17, 27, 30, 33
			2	2, 10, 14, 15, 16, 18, 19, 22, 25, 28, 34
18	x_6, x_8, x_9, x_{12}, x_{17}	2	1	3, 5, 8, 11, 12, 13, 15, 22, 23, 25, 26, 27, 29, 32, 35
			2	1, 7, 10, 14, 16, 17, 18, 20, 28, 34
19	x_6, x_8, x_9, x_{12}, x_{17}	2	1	4, 6, 11, 13, 22, 23, 25, 26, 27, 29, 30, 32, 33, 35
			2	2, 9, 10, 14, 16, 17, 18, 19, 20, 24, 28, 31, 34
20	x_6, x_8, x_9, x_{12}, x_{17}	2	1	9, 10, 14, 16, 17, 18, 19, 20, 24, 28, 31, 34
			2	11, 12, 13, 15, 22, 23, 25, 26, 29, 30, 32, 33, 35