МЕЖГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«БЕЛОРУССКО-РОССИЙСКИЙ УНИВЕРСИТЕТ»

Кафедра «Автоматизированные системы управления»

ФАКТОРНЫЙ И КОМПОНЕНТНЫЙ АНАЛИЗ

Методические рекомендации к лабораторным работам для магистрантов специальности 7-06-0612-03 «Системы управления информацией» очной и заочной форм обучения

Могилев 2025

УДК 519.237
ББК 22.172
Ф18

Рекомендовано к изданию
учебно-методическим отделом
Белорусско-Российского университета

Одобрено кафедрой «Автоматизированные системы управления» «23» сентября 2025 г., протокол № 2

Составители: канд. физ.-мат. наук, доц. В. А. Ливинская;
ст. преподаватель И. А. Беккер

Рецензент канд. техн. наук, доц. С. К. Крутолевич

Методические рекомендации к лабораторным работам содержат краткие теоретические сведения, задания, общие требования к отчету. Предназначены для магистрантов специальности 7-06-0612-03 «Системы управления информацией» очной и заочной форм обучения.

Учебное издание

ФАКТОРНЫЙ И КОМПОНЕНТНЫЙ АНАЛИЗ

Ответственный за выпуск А. И. Якимов

Корректор А. А. Подошевко

Компьютерная верстка Н. П. Полевничая

Подписано в печать . Формат 60×84/16. Бумага офсетная. Гарнитура Таймс.
Печать трафаретная. Усл. печ. л. . Уч.- изд. л. . Тираж 21 экз. Заказ №

Издатель и полиграфическое исполнение:
Межгосударственное образовательное учреждение высшего образования
«Белорусско-Российский университет».
Свидетельство о государственной регистрации издателя,
изготовителя, распространителя печатных изданий
№ 1/156 от 07.03.2019.
Пр-т Мира, 43, 212022, г. Могилев.

Содержание

Введение

Изучение дисциплины «Факторный и компонентный анализ» дает магистрантам практические умения и навыки анализа данных, которые будут полезными при статистической обработке информации, написании диссертационных работ.

Общий порядок выполнения работ следующий:

1) изучить теоретические сведения по теме работы;
2) получить задание у преподавателя;
3) реализовать задание;
4) дать обоснование полученного решения;
5) сделать выводы по результатам исследований;
6) оформить отчет в печатном виде.

Отчет должен удовлетворять определенным требованиям и содержать следующие части:

1) цель работы;
2) постановка задачи;
3) ход и результаты выполнения задания, с необходимыми скриншотами;
4) выводы по итогам выполнения задания.

Лабораторная работа № 1. Изучение возможностей применения встроенных функций EXCEL для решения задач, связанных со статистической обработкой информации

Цель работы: изучить возможности встроенных средств EXCEL для статистического анализа данных.

Порядок выполнения работы

1 Изучить теоретические сведения [1–3].
2 Получить задание у преподавателя.
3 Реализовать задание.
4 Дать обоснование полученного решения.
5 Сделать выводы по результатам исследований и оформить отчет.

Задание

1 Сгенерировать генеральную совокупность объемом N с заданным законом распределения и случайным образом сформировать выборочную совокупность объемом n (согласно своему варианту, полученному у преподавателя).

2 Вычислить описательную статистику для выборки и генеральной совокупности.

3 Построить гистограмму для выборки и проверить гипотезу о виде закона распределения, используя критерий Пирсона при уровне значимости α.

В отчете отразить:

1) гистограмму распределения;
2) выбранную гипотезу о виде закона распределения;
3) вычисленное значение критерия;
4) критическое значение;
5) вывод о принятии или непринятии гипотезы.

Методические указания

Основными средствами анализа статистических данных в EXCEL являются статистические процедуры надстройки Пакет анализа и статистические функции из библиотеки встроенных функций.

В работе нужно использовать такие статистические процедуры надстройки «Анализ данных» пакета анализа EXCEL, как Генерация случайных чисел; Выборка; Описательная статистика.

При анализе вариационных рядов распределения большое значение имеет, насколько эмпирическое распределение признака соответствует нормальному. Для этого частоты фактического распределения нужно сравнить с теоретическими, которые характерны для нормального распределения. Значит, нужно по фактическим данным вычислить теоретические частоты кривой нормального

распределения, являющиеся функцией нормированных отклонений.

Иначе говоря, эмпирическую кривую распределения нужно выровнять кривой нормального распределения. Объективная характеристика соответствия теоретических и эмпирических частот может быть получена при помощи специальных статистических показателей, которые называют критериями согласия.

Критерием согласия называют критерий, который позволяет установить, является ли расхождение эмпирического и теоретического распределений случайным или значимым, т. е. согласуются ли данные наблюдений с выдвинутой статистической гипотезой или не согласуются. Распределение генеральной совокупности, которое она имеет в силу выдвинутой гипотезы, называют теоретическим.

Возникает необходимость установить критерий, который позволял бы судить, является ли расхождение между эмпирическим и теоретическим распределениями случайным или значимым. Если расхождение окажется случайным, то считают, что данные наблюдений (выборки) согласуются с выдвинутой гипотезой о законе распределения генеральной совокупности и, следовательно, гипотезу принимают; если же расхождение окажется значимым, то данные наблюдений не согласуются с гипотезой и ее отвергают.

Обычно эмпирические и теоретические частоты различаются в силу того, что расхождение случайно и связано с ограниченным количеством наблюдений; расхождение неслучайно и объясняется тем, что статистическая гипотеза о том, что генеральная совокупность распределена нормально, – ошибочна.

Таким образом, критерии согласия позволяют отвергнуть или подтвердить правильность выдвинутой при выравнивании ряда гипотезы о характере распределения в эмпирическом ряду.

Эмпирические частоты получают в результате наблюдения. Теоретические частоты рассчитывают по формулам согласно функции распределения предполагаемого закона.

Для проверки соответствия выборочных данных предполагаемому закону распределения необходимо построить гистограмму и получить числовые характеристики выборки.

Для проверки гипотезы с помощью критерия Пирсона в EXCEL следует воспользоваться [3].

Контрольные вопросы

1 Какие функции EXCEL применяются для получения числовых характеристик выборки?

2 На основании чего выдвигается гипотеза о законе распределения?

3 Как описывается закон распределения в Вашем случае?

4 Какой критерий для проверки гипотезы использовался?

Лабораторная работа № 2. Первичная обработка опытных данных при помощи модуля Basic Statistics/Tables в ППП STATISTICA

Цель работы: изучить возможность модуля Basic Statistics/Tables в ППП STATISTICA при первичном анализе статистической информации.

Порядок выполнения работы

1 Изучить теоретические сведения по теме «Статистические методы анализа данных. Проверка статистических гипотез».

2 Получить задание у преподавателя.

3 Реализовать задание.

4 Дать обоснование полученного решения.

5 Сделать выводы по результатам исследований и оформить отчет.

Задание

На основании двух выборок, полученных у преподавателя (файл Туризм.xls), согласно своему варианту проанализировать данные (таблица 1).

Таблица 1 – Исходные данные

Вариант	Регион 1	Регион 2	Признак
1	Западная Европа (код 1)	Восточная Европа(код 2)	Размер расходов туристов в течение проживания за день
2	Западная Европа (код 1)	Восточная Европа(код 2)	Доход от туризма
3	Западная Европа (код 1)	Восточная Европа(код 2)	Средний месячный доход туристов
4	Западная Европа (код 1)	Скандинавские страны (код 3)	Количество туристов
5	Западная Европа (код 1)	Скандинавские страны (код 3)	Размер расходов в течение проживания за день
6	Западная Европа (код 1)	Скандинавские страны (код 3)	Количество экскурсий
7	Восточная Европа (код 2)	Скандинавские страны (код 3)	Размер расходов в течение всего периода проживания
8	Восточная Европа (код 2)	Скандинавские страны (код 3)	Объем чаевых
9	Восточная Европа (код 2)	Скандинавские страны (код 3)	Траты туристов на билеты

1 Определить числовые характеристики выборок (с помощью команды Описательные статистики модуля Basic Statistics/Tables).

2 Проверить гипотезу о согласии с нормальным распределением данных выборочных совокупностей, используя критерии Колмогорова – Смирно-

ва (K-S), Лиллиефорса, Шапиро – Уилка при построении гистограмм.

3 Проверить гипотезу о равенстве средних в генеральных совокупностях (при условии гомогенности дисперсий) с помощью *t*-критерия и с помощью доверительных интервалов.

4 Подтвердить выводы п. 3 с помощью диаграммы 2D box-plot.

Методические указания

В модуле Basic Statistics/Tables в ППП STATISTICA (рисунок 1) реализована возможность проверки гипотез о равенстве выборочного среднего некоторому заданному числу (*t*-критерий для одной выборки), а также *t*-критерий для двух независимых выборок (двухвыборочный *t*-критерий), который проверяет гипотезу о равенстве средних в двух выборках (предполагается нормальность распределения переменных, а также равенство дисперсий выборок).

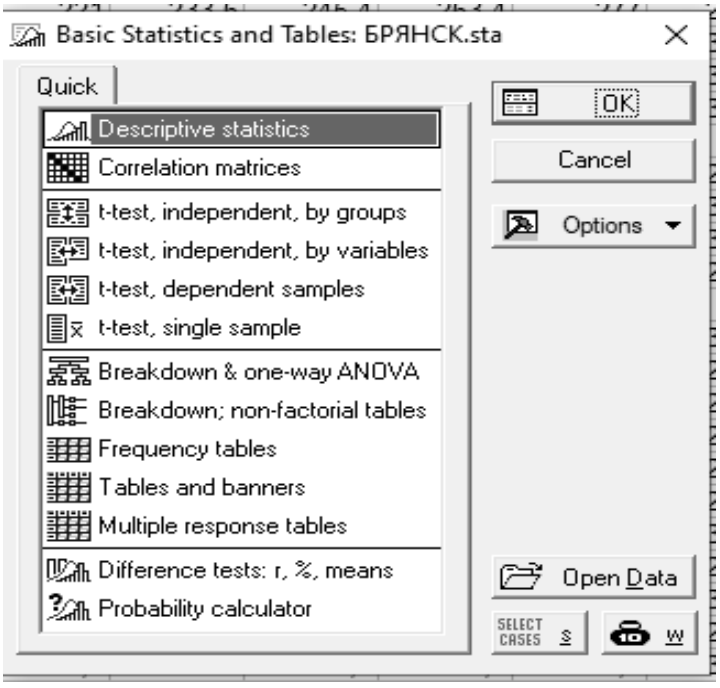

Рисунок 1 – Модуль проверки статистических гипотез

Задание для самостоятельной работы

На основании данных из открытых источников сформулировать и проверить гипотезы о статистически значимых различиях в выборках, распределение которых отлично от нормального.

Результат исследования оформить в виде выступления на научной конференции.

Контрольные вопросы

1 Какие гипотезы относятся к параметрическим?

2 Опишите алгоритм проверки гипотезы о равенстве дисперсий.

3 Опишите алгоритм проверки гипотезы о равенстве математических ожиданий двух совокупностей.

4 Что такое ошибка первого рода при проверке статистической гипотезы?

Лабораторная работа № 3. Корреляционный анализ количественных и номинальных данных в ППП STATISTICA

Цель работы: изучить возможности ППП STATISTICA при установлении зависимости между переменными и оценке характера зависимости.

Порядок выполнения работы

1 Изучить теоретический материал по теме «Основные задачи интеллектуального анализа данных. Корреляционный анализ».

2 Получить задание у преподавателя.

3 Реализовать задание.

4 Дать обоснование полученного решения.

5 Сделать выводы по результатам исследований и оформить отчет.

Задание

Анализируются данные, являющиеся случайной выборкой из записей о продажах домов в определенный период.

Необходимо:

1) выполнить проверку количественных данных на принадлежность выборок генеральным совокупностям, имеющим нормальное распределение;

2) определить силу, направление и статистическую достоверность связи между количественными данными, распределенными по нормальному закону, с помощью коэффициента линейной корреляции Пирсона;

3) определить силу, направление и статистическую достоверность связи между количественными данными, распределение которых не подчиняется нормальному закону распределения.

Методические указания

Для определения степени тесноты линейной зависимости между признаками, имеющими нормальное распределение, в многомерном статистическом анализе используется корреляционная матрица, парные коэффициенты корреляции r_{ij} между i и j признаками могут быть определены в модуле Basic Statistics and table.

В случае, если гипотеза о нормальности выборок отвергается либо данных недостаточно, используются ранговые коэффициенты корреляции Спирмена и Кэнделла, они могут быть найдены в модуле Nonparametric Statistics (рисунок 2).

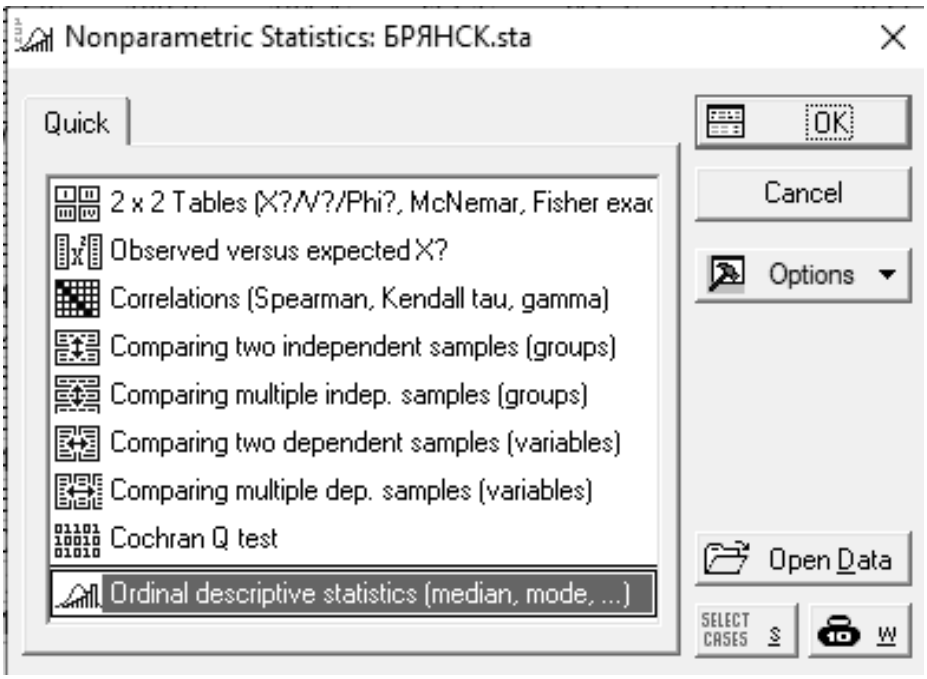

Рисунок 2 – Модуль Непараметрические статистики

Задание для самостоятельной работы

На основании данных из открытых источников сформулировать и проверить гипотезы о наличии статистически значимой связи между количественными признаками. Результат исследования оформить в виде выступления на научной конференции.

Контрольные вопросы

1 Что такое корреляция признаков?

2 Как графически определить наличие/отсутствие связи между признаками?

3 Тесноту какого типа связи можно оценить с помощью коэффициента корреляции Пирсона?

4 Каковы назначение, область применения и ограничения ранговых коэффициентов корреляции?

5 В каких пределах может меняться значение коэффициентов корреляции Спирмена и Кэндалла?

6 На основании чего делается вывод о достоверности статистической связи?

Лабораторная работа № 4. Регрессионный анализ количественных и номинальных данных в ППП STATISTICA

Цель работы: изучить возможности ППП STATISTICA при проведении множественного регрессионного анализа с количественными и номинальными переменными.

Порядок выполнения работы

1 Изучить теоретический материал по теме «Основные задачи интеллектуального анализа данных. Регрессионный анализ».

2 Получить задание у преподавателя.

3 Реализовать задание.

4 Дать обоснование полученного решения.

5 Сделать выводы по результатам исследований.

6 Оформить отчет.

Задание

1 По данным, полученным в лабораторной работе № 3, отобрать факторы, оказывающие наибольшее влияние на результативный признак.

2 Оценить коэффициенты множественной линейной регрессии в модуле Multiple Regression ППП STATISTICA [6].

3 Проверить гипотезу о статистической значимости оценок параметров модели на основе *F*- и *t*-критериев.

4 Оценить наличие гетероскедастичности в остатках.

5 Построить доверительные интервалы для значимых оценок параметров модели.

6 Осуществить точечный прогноз индивидуального значения показателя.

7 Построить доверительный интервал для прогноза индивидуального значения показателя.

Методические указания

Пусть вектор Y (результативный признак) зависит от k факторных признаков, представленных матрицей X.

Зависимость при линейной форме связи в матричном виде имеет вид

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_k x_k,$$

где x_i – известные значения факторных признаков (столбцы матрицы X);

α_i – неизвестные коэффициенты, подлежащие определению.

Для оценки вектора параметров множественной линейной регрессионной модели служит метод наименьших квадратов (МНК).

Проверка качества уравнения регрессии заключается в следующих действиях:

1) проверка значимости всех α_j;

2) проверка общего качества уравнения регрессии с помощью коэффициента множественной детерминации R^2;

3) проверка свойств данных, выполнение которых предполагалось при оценивании уравнений.

Расчеты выполняются в модуле Multiply Regression программы STATISTICA (рисунок 3).

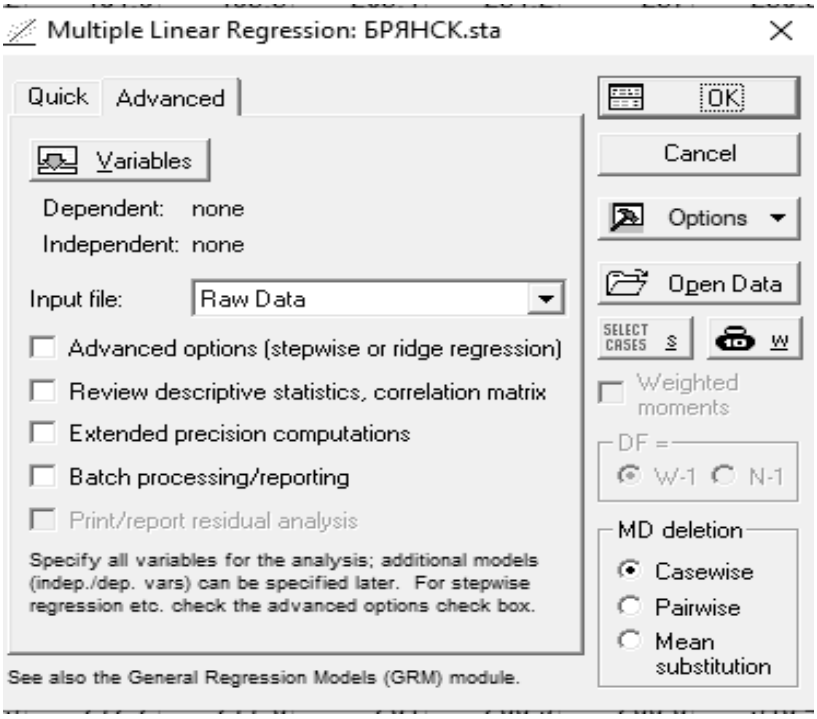

Рисунок 3 – Модуль Множественная регрессия

Задание для самостоятельной работы

1 На основании данных из открытых источников оценить модель линейной регрессии между количественными признаками.

2 Результат исследования оформить в виде выступления на научной конференции.

Контрольные вопросы

1 Как оценивается модель множественной регрессии в матричном виде?

2 Как проверяется качество регрессионной модели?

3 Как интерпретируются коэффициенты уравнения множественной регрессии?

4 Какими методами можно обнаружить гетероскедастичность?

Лабораторная работа № 5. Кластерный анализ в ППП STATISTICA

Цель работы: научиться методам группирования многомерных данных; показать возможности визуализации последовательного формирования кластеров сходных объектов.

Порядок выполнения работы

1 Изучить теоретический материал по теме «Кластерный анализ».
2 Получить задание у преподавателя.
3 Реализовать задание.
4 Дать обоснование полученного решения.
5 Сделать выводы по результатам исследований.
6 Оформить отчет.

Задание

Имеются данные о персонах (работниках, студентах, клиентах, гражданах).

Требуется провести кластерный анализ с целью обнаружения групп кандидатов, близких по своим качествам. Сравнить решения, полученные при разбиении на три и четыре кластера, с применением программы STATISTICA провести классификацию объектов. Сравнить решения, полученные методом иерархического кластерного анализа и методом *k*-средних.

В каждом случае указать:

– методику вычисления основных видов расстояний между объектами и между кластерами;
– объекты, вошедшие в каждый кластер;
– описательные статистики каждого кластера;
– график средних;
– обоснование разбиения на кластеры;
– признак, по которому кластеры различаются наибольшим образом.

Методические указания

Методы кластерного анализа позволяют выделить из исследуемой совокупности объектов *кластеры* – скопления объектов с близкими значениями параметров.

Методы кластерного анализа можно разделить на две большие категории по алгоритму действия. Первая группа методов называется *иерархическими*, т. к. в процессе работы метода строится иерархия вложенности кластеров, обычно представляемая на графике – *дендрограмме*. На каждом шаге агломеративной иерархической процедуры объединяется пара ближайших кластеров.

Методы второй категории называются *итерационными*, т. к. они основаны на поиске оптимального положения центров кластеров на каждой итерации –

последовательного рассмотрения всех объектов исходной выборки.

Среди итерационных методов наиболее распространённым является *метод k-средних*. На первом его шаге необходимо задать требуемое количество кластеров k и начальные центры их тяжести. В качестве этих начальных центров обычно используются первые k наблюдений выборки, однако в некоторых случаях это может привести к недостаточному качеству полученного решения. Поэтому возможно применить иерархическую процедуру на случайной выборке и затем использовать полученные центры в итерационной процедуре.

Все методы кластерного анализа реализованы в модуле Clustering Method в программе STATISTICA (рисунок 4).

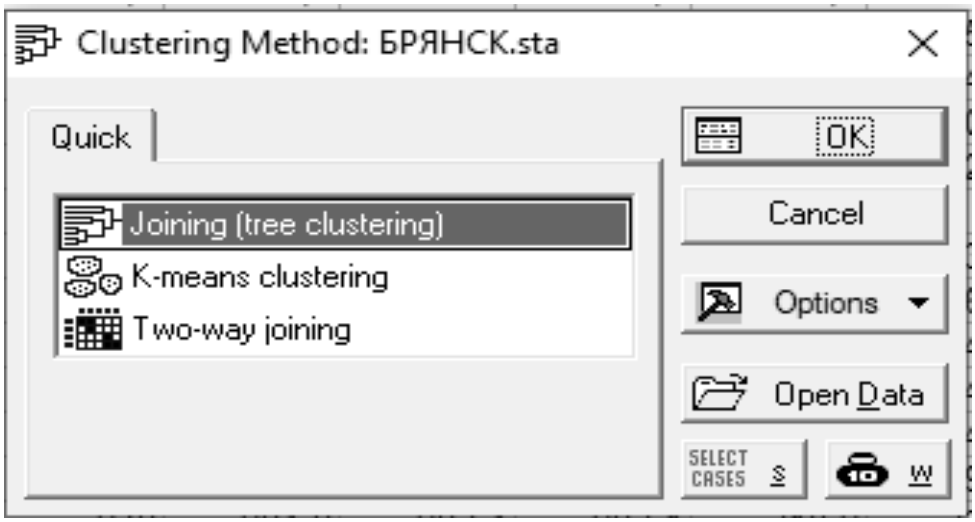

Рисунок 4 – Модуль Кластерный анализ

Расстояние между кластерами определяется с помощью следующих основных методов:

– связь между группами – расстояние между двумя кластерами определяется как среднее значение расстояний между всеми возможными парами наблюдений, причём одно наблюдение берётся из одного кластера, а второе – из другого;

– связь внутри групп – расстояние между двумя кластерами определяется как среднее значение расстояний между всеми возможными парами наблюдений из обоих кластеров, включая пары наблюдений внутри кластеров;

– ближний сосед – расстояние между двумя кластерами определяется как минимальное из всех расстояний между всеми возможными парами наблюдений из разных кластеров;

– дальний сосед – расстояние между двумя кластерами определяется как максимальное из всех расстояний между всеми возможными парами наблюдений из разных кластеров;

– центроидная кластеризация – расстояние между двумя кластерами определяется как расстояние между центрами тяжести обоих кластеров;

– медианная кластеризация – расстояние между двумя кластерами определяется как взвешенное центроидное расстояние между кластерами, где веса

соответствуют размеру каждого кластера;

– метод Варда – в этом методе объединяются только те два кластера, для которых прирост внутрикластерной дисперсии минимален.

Наиболее универсальными методами являются метод Варда и метод межгрупповой связи.

Задание для самостоятельной работы

На основании данных из открытых источников выполнить разбиение на группы объектов, характеризующихся набором признаков, с помощью кластерного анализа. Результат исследования оформить в виде выступления на научной конференции.

Контрольные вопросы

1 Как определяется в кластерном анализе мера близости объектов?

2 Как определяется расстояние между кластерами?

3 Какие методы кластерного анализа относятся к иерархическим агломеративным методам?

4 В чем суть итеративных методов? Опишите метод *k*-средних.

5 Что такое функционалы качества разбиения? Приведите примеры функционалов разбиения при известном числе кластеров и неизвестном числе кластеров.

6 Какие статистические критерии используются для проверки значимости различия кластеров?

Лабораторная работа № 6. Логистическая регрессия как метод классификации в ППП STATISTICA

Цель работы: изучить алгоритм классификации объектов с помощью логистической регрессии и применить его на практике.

Порядок выполнения работы

1 Изучить теоретический материал по теме «Логистическая регрессия».

2 Получить задание у преподавателя.

3 Реализовать задание.

4 Дать обоснование полученного решения.

5 Сделать выводы по результатам исследований и оформить отчет.

Задание

Кредитным отделом банка накоплена информация о кредитополучателях и их платёжной дисциплине. Информация о клиентах содержит сведения о их возрасте, образовании, месте работы, зарплате, наличии недвижимого имущества и другие сведения, а также их кредитную историю. Одной из

характеристик является наличие или отсутствие пени за просрочку платежей.

Требуется построить логистическую классификационную модель, которая будет определять, будет ли новый клиент нарушать график платежей или нет. Сравнить несколько вариантов логит-модели с различными регрессионными уравнениями.

Методические указания

Логистическая регрессия применяется для прогнозирования вероятности возникновения некоторого события по значениям множества признаков. Для этого вводится так называемая зависимая переменная y, принимающая лишь одно из двух значений – как правило, это числа 0 (событие не произошло) и 1 (событие произошло), и множество независимых переменных (также называемых признаками, предикторами или регрессорами) – вещественных x_1, x_2, …, x_n, на основе значений которых требуется вычислить вероятность принятия того или иного значения зависимой переменной.

Эта модель часто применяется для решения задач классификации. Подбор коэффициентов регрессионного уравнения осуществляется на обучающей выборке с помощью метода максимального правдоподобия в модуле Nonlinear Estimation-Quick Logit regression в программе STATISTICA.

Для оценки качества классификатора используется таблица 2.

Таблица 2 – Таблица сопряженности

Решение по тестируемомуметоду	Фактическое состояние объектов	
	1	0
1	*TR*	*FP*
0	*FN*	*TN*

По таблице 3 рассчитываются такие важные характеристики классификатора, как чувствительность и специфичность.

Чувствительность (Sensitivity) (доля истинно положительных случаев, которые были правильно идентифицированы тестируемым методом) по формуле

$$Se = TP/(TP + FN) \cdot 100\ \%.$$

Специфичность (Specificity) – доля истинно отрицательных случаев – по формуле

$$Sp = TN/(TN + FP) \cdot 100\ \%.$$

Модель с высокой чувствительностью часто дает истинный результат при наличии положительного исхода (обнаруживает положительные примеры).

Наоборот, модель с высокой специфичностью чаще дает истинный результат при наличии отрицательного исхода (обнаруживает отрицательные примеры).

Графической интерпретацией результатов классификации с помощью ло-

гистической регрессии служит ROC-кривая, а площадь под ней (AUC) является показателем качества классификационной модели (таблица 3).

Таблица 3 – Соответствие значения AUC и качества логистической модели

Интервал значений AUC	Качество модели
0,9...1,0	Отличное
0,8...0,9	Очень хорошее
0,7...0,8	Хорошее
0,6...0,7	Среднее
0,5...0,6	Неудовлетворительное

Задание для самостоятельной работы

На основании данных из открытых источников построить правило классификации объектов, характеризующихся набором признаков, с помощью логистической регрессионной модели. Оценить качество классификатора, построить ROC-кривую. Результат исследования оформить в виде выступления на научной конференции.

Контрольные вопросы

1 Для чего используется логистическая модель?
2 Какие показатели рассчитываются на основе матрицы сопряжённости?
3 Что показывает показатель AUC?
4 В каких диапазонах может изменяться AUC?

Лабораторная работа № 7. Классификация многомерных наблюдений с обучением в ППП STATISTICA

Цель работы: изучить основные процедуры дискриминантного анализа – дискриминации и классификации; получить навыки построения и определения количества дискриминантных функций и их разделительной способности, нахождения классифицирующих функций с использованием функций Фишера и расстояния Махаланобиса.

Порядок выполнения работы

1 Изучить теоретический материал по теме «Классификация и распознавание образов. Логистическая регрессия».
2 Получить задание у преподавателя.
3 Реализовать задание.
4 Дать обоснование полученного решения.
5 Сделать выводы по результатам исследований.
6 Оформить отчет.

Задание

1 Изучить теоретический материал по теме «Дискриминантный анализ».

2 В файле данных Heart.sta приведены данные о возрасте, давлении, весе, росте и уровне холестерина пациентов 1950 г., а также их состояние на 1968 г. (переменная DTH принимает значение 1, если пациент умер, 0 – если нет). Провести дискриминантный анализ для проверки возможности прогнозирования летального исхода на основании давления, холестерина и физических данных пациентов.

3 Войти в пакет STATISTICA (модуль Discriminant analysis).

4 Ввести исходные данные для проведения дискриминантного анализа в рабочий файл.

5 Проверить межгрупповые и общие корреляции и ковариации. Определить значения средних и стандартных девиаций по группам и выяснить, для какой из переменных сильнее отличаются значения средних.

6 Провести пошаговый анализ. Выяснить, какая из переменных лучше всего подходит для дискриминации. Проверить значения толерантности.

Методические указания

Дискриминантный анализ – раздел многомерного статистического анализа, который позволяет предсказать принадлежность объектов к двум или более непересекающимся группам. Исходными данными для дискриминантного анализа является множество объектов, разделенных на группы так, что каждый объект может быть отнесен только к одной группе. Для каждого из объектов имеются данные по ряду количественных переменных. Такие переменные называются дискриминантными переменными или предикторами.

Задачами дискриминантного анализа является определение:

– решающих правил, позволяющих по значениям дискриминантных переменных (предикторов) отнести каждый объект к одной из известных групп;

– «веса» каждой дискриминантной переменной для разделения объектов на группы.

Ядром дискриминантного анализа является построение так называемой дискриминантной функции

$$d = b_1 x_1 + b_2 x_2 + ... + b_n x_n + a,$$

где x_i ($i = 1, \ldots, n$) – значения переменных, соответствующих рассматриваемым случаям;

b_i, a – коэффициенты, которые и предстоит оценить с помощью дискриминантного анализа.

Необходимо определить такие коэффициенты, чтобы по значениям дискриминантной функции можно было с максимальной четкостью провести разделение по группам.

Признак считается влияющим на разделение по группам, если отношение

внутригрупповой суммы квадратов к общей сумме квадратов близко к нулю (с учетом значения статистики F и соответствующего уровня значимости p).

Виды дискриминантного анализа

Пошаговый анализ с включением. В пошаговом анализе дискриминантных функций модель дискриминации строится по шагам. Точнее, на каждом шаге просматриваются все переменные и находится та из них, которая вносит наибольший вклад в различие между совокупностями. Эта переменная должна быть включена в модель на данном шаге, и происходит переход к следующему шагу.

Пошаговый анализ с исключением. Можно также двигаться в обратном направлении. В этом случае все переменные будут сначала включены в модель, а затем на каждом шаге будут устраняться переменные, вносящие малый вклад в предсказания.

Тогда в качестве результата успешного анализа можно сохранить только «важные» переменные в модели, т. е. те переменные, чей вкладв дискриминацию больше остальных. Эта пошаговая процедура руководствуется соответствующим значением F для включения и соответствующим значением F для исключения. Значение F-статистики для переменной указывает на ее статистическую значимость при дискриминации между совокупностями, т. е. она является мерой вклада переменной в предсказание членства в совокупности.

Процедура дискриминантного анализа состоит из следующих стадий.

1 Разделение выборки на две части, т. е. формирование анализируемой выборки (части общей выборки, которую используют для вычисления дискриминантной функции) и тестируемой выборки (части общей выборки, которую используют для проверки результатов расчета на основании анализируемой выборки).

2 Выбор переменных – предикторов.

На начальном этапе дискриминантного анализа для предикторов формируется корреляционная матрица.

В данном контексте она имеет особый смысл, называется общей внутригрупповой корреляционной матрицей и содержит средние коэффициенты корреляции для двух или более корреляционных матриц (каждая для одной группы).

3 Вычисление параметров дискриминантной функции.

Решение поставленной задачи осуществляется в модуле Discriminant Function Analysis программы STATISTICA.

4 Интерпретация результатов.

Контрольные вопросы

1 В чем отличие дискриминантного анализа от кластерного?
2 Для чего строится каноническая дискриминантная функция?
3 Как определяются коэффициенты дискриминантной функции?
4 Что такое константа дискриминации?

Лабораторная работа № 8. Факторный анализ и его реализация в ППП STATISTICA

Цель работы: овладеть и научиться практически применять знания и умения в представлении эмпирических переменных в качестве линейных комбинаций меньшего числа некоторых других переменных.

Порядок выполнения работы

1 Получить задание у преподавателя.
2 Реализовать задание.
3 Дать обоснование полученного решения.
4 Сделать выводы по результатам исследований.
5 Оформить отчет.

Задание

1 Изучить теоретический материал по темам «Факторный анализ как метод редукции данных», «Измерение и оценка факторов», «Применение факторного анализа к задачам классификации».

2 Войти в пакет STATISTICA (модуль Factor analysis).

3 Открыть файл с набором данных, описывающим показатели финансовой структуры банков (файл данных Bank.sta).

4 Провести факторный анализ по переменным AGE–USTAV, определить оптимальное количество факторов и интерпретировать их:

– провести корреляционный анализ и рассчитать описательную статистику для факторов;

– определить собственные значения корреляционной матрицы, на основании которых сделать вывод о количестве факторов, наиболее полно описывающих банки с различными финансовыми показателями;

– с помощью вращения факторов подобрать оптимальное количество факторов;

– сохранить факторные значения в каком-либо файле с переменной ИСХОД. Построить диаграмму рассеяния в пространстве полученных факторов для «лопнувших» и еще действующих банков. Сравнить «лопнувшие» и действующие банки по факторным значениям.

Методические указания

Пусть проводится p наблюдений над n признаками X_1, X_2, ..., X_n. Под наблюдениями понимаем набор из p однотипных объектов, для каждого из которых фиксируются значения заданного набора из n признаков.

Таким образом, исходными данными служит набор из n p-мерных векторов. При этом предполагается, что все данные подвергнуты нормированию и центрированию.

Основным предположением линейной модели факторного анализа является

предположение о том, что признаки выражаются через факторы линейно.

Существует несколько методов решения задачи факторного анализа. Однако в большинстве практических исследований применяется метод главных компонент.

Идея метода главных компонент заключается в поиске ортогональной системы из n векторов со специальными свойствами.

Первоначально модель содержит такое же число факторов F_k (главных компонент), что и косвенных признаков. Это позволяет отказаться от введения специфических факторов U_i.

Таким образом, в этой новой системе координат ковариационная матрица должна иметь диагональную форму (ограничиваемся здесь случаем, когда все собственные значения матрицы ковариаций простые).

Находятся собственные значения λ_1, λ_2, …, λ_n корреляционной матрицы, являющиеся дисперсиями новых факторов. С помощью метода каменистой сыпи отбирают факторы, обеспечивающие более 70 % кумулятивной дисперсии.

Далее вычисляют коэффициенты корреляции между главными факторами и исходными признаками и с их помощью получают координаты объектов в новой системе главных факторов.

Данный метод позволяет снизить количество факторов, описывающих совокупность объектов.

В случае необходимости дают смысловую интерпретацию факторам, но чаще их используют для проведения дальнейшего анализа (кластерного, регрессионного).

Контрольные вопросы

1 В чем суть факторного анализа?

2 Опишите идею метода главных компонент.

3 Как производится отбор факторов, описывающих данные наиболее оптимальным образом?

4 Как оценивается значимость модели факторного анализа?

5 Для чего в факторном анализе используют процедуру вращения факторов?

Лабораторная работа № 9. Компонентный анализ и его реализация в ППП STATISTICA

Цель работы: изучить особенности применения компонентного анализа в среде STATISTICA для изучения структуры зависимости в данных и извлечения знаний.

Порядок выполнения работы

1 Изучить теоретический материал по темам «Сущность метода главных компонент», «Применение метода главных компонентов в задачах

распознавания».

2 Получить задание у преподавателя.

3 Реализовать задание.

4 Дать обоснование полученного решения.

5 Сделать выводы по результатам исследований.

6 Оформить отчет.

Задание

1 Открыть файл *Swiss Fertility.xls*, в котором рассматривается выборка – 47 франкоговорящих провинций Швейцарии в 1888 г. В набор данных вошли показатели социального и экономического развития.

Все переменные принимают значения в интервале [0, 100].

2 По представленным выборочным данным провести компонентный анализ (с применением программы STATISTICA), позволяющий построить обобщенные характеристики, описывающие различия в социально-экономической ситуации в провинциях Швейцарии:

− рассчитать выборочные характеристики;

− нормировать данные;

− рассчитать матрицы собственных значений и собственных векторов;

− рассчитать матрицы факторных нагрузок и значений главных компонент;

− ранжировать регионы внутри по первой главной компоненте.

3 Построить уравнение зависимости рождаемости от главных компонент.

Контрольные вопросы

1 Для чего используется метод главных компонент?

2 Что такое корреляционная матрица?

3 Как вычислить собственные числа и собственные векторы корреляционной матрицы?

4 По каким формулам вычисляются оценки среднего значения, дисперсии и коэффициентов корреляции?

5 Какие критерии используются для оценки результатов метода?

Список литературы

1 **Кулешова, О. В.** Microsoft Excel 2016/2013. Расширенные возможности. Решение практических задач / О. В. Кулешова. – М. : Специалист, 2016. – 100 с.

2 **Мхитарян, В. С.** Анализ данных в MS Excel : учеб. пособие / В. С. Мхитарян, В. Ф. Шишов, А. Ю. Козлов. – М. : КУРС, 2019. – 368 с.

3 **Дубров, А. М.** Многомерные статистические методы : учебник / А. М. Дубров, В. С. Мхитарян, Л. И. Трошин. – М. : Финансы и статистика, 2011. – 310 с.