МЕЖГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«БЕЛОРУССКО-РОССИЙСКИЙ УНИВЕРСИТЕТ»

Кафедра «Автоматизированные системы управления»

ФАКТОРНЫЙ И КОМПОНЕНТНЫЙ АНАЛИЗ

Методические рекомендации к лабораторным работам для магистров специальности 1-40 80 02 «Системный анализ, управление и обработка информации» очной и заочной форм обучения

Могилев 2020

УДК 519.237
ББК 22.172
Ф18

Рекомендовано к изданию
учебно-методическим отделом
Белорусско-Российского университета

Одобрено кафедрой «Автоматизированные системы управления» «17» марта 2020 г., протокол № 8

Составитель канд. физ.-мат. наук, доц. В. А. Ливинская

Рецензент канд. техн. наук, доц. И. В. Лесковец

В методических рекомендациях представлены задания, методические указания, контрольные вопросы к каждой лабораторной работе.

Учебно-методическое издание

ФАКТОРНЫЙ И КОМПОНЕНТНЫЙ АНАЛИЗ

Ответственный за выпуск	А. И. Якимов
Корректор	И. В. Голубцова
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 16 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, Могилев.

Содержание

Введение

Развитие информационных технологий последние десятилетия предъявляет к соискателям наличие компетенций в различных предметных областях. Так, появилась новая специализация – аналитик данных, компетенции которой, зачастую, необходимо осваивать самостоятельно. Аналитик данных извлекает из данных смысл: структурирует их, формулирует и проверяет гипотезы, находит закономерности и делает выводы. Его работа помогает принимать решения в бизнесе, управлении и науке.

Методические рекомендации предназначены для получения практических навыков магистрантами по анализу информации (большие массивы данных), представленной в открытом доступе. Могут быть полезными при написании диссертационных работ и проведении различного рода исследований, связанных со статистической обработкой информации.

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

Цель работы: научиться пользоваться встроенными средствами Excel для статистического анализа данных.

Порядок выполнения работы

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

Требования к отчету

1 Цель работы.
2 Постановка задачи.
3 Результаты выполнения задания.
4 Выводы.

Задание

1 Сгенерировать генеральную совокупность объемом N с заданным законом распределения и случайным образом сформировать выборочную 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 (Analysis ToolPak) и статистические функции библиотеки встроенных функций. Основные сведения обо всех этих средствах имеются в электронной справочной системе Excel.

В работе использовать следующие статистические процедуры надстройки «Анализ данных» пакета анализа EXCEL:

– генерация случайных чисел;

– выборка;

– гистограмма;

– 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 (правило), 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 воспользоваться [3, 4].

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.

6 Оформить отчет.

Требования к отчету

1 Цель работы.

2 Постановка задачи.

3 Результаты выполнения задания.

4 Выводы.

Задание

На основании двух выборок, полученных у преподавателя (файл Туристские поездки.xls), согласно своему варианту (таблица 1):

1) определить числовые характеристики выборок с помощью команды «Описательные статистики (Descriptive statistics)» модуля Basic Statistics/Tables [6];

2) проверить гипотезу о согласии с нормальным распределением данных выборочных совокупностей, используя критерии Колмогорова–Смирнова (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];

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

Таблица 1 – Исходные данные

Вариант	Регион 1	Регион 2	Признак
1	Западная Европа (код 1)	Восточная Европа (код 2)	Размер дополнительных расходов в течение проживания за день
2	Западная Европа (код 1)	Восточная Европа (код 2)	Объем чаевых за одно проживание
3	Западная Европа (код 1)	Восточная Европа (код 2)	Средний месячный доход туристов
4	Западная Европа (код 1)	Скандинавские страны (код 3)	Средний месячный доход туристов
5	Западная Европа (код 1)	Скандинавские страны (код 3)	Размер дополнительных расходов в течение проживания за день
6	Западная Европа (код 1)	Скандинавские страны (код 3)	Объем чаевых за одно проживание
7	Восточная Европа (код 2)	Скандинавские страны (код 3)	Размер дополнительных расходов в течение проживания за день
8	Восточная Европа (код 2)	Скандинавские страны (код 3)	Объем чаевых за одно проживание
9	Восточная Европа (код 2)	Скандинавские страны (код 3)	Средний месячный доход туристов

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).

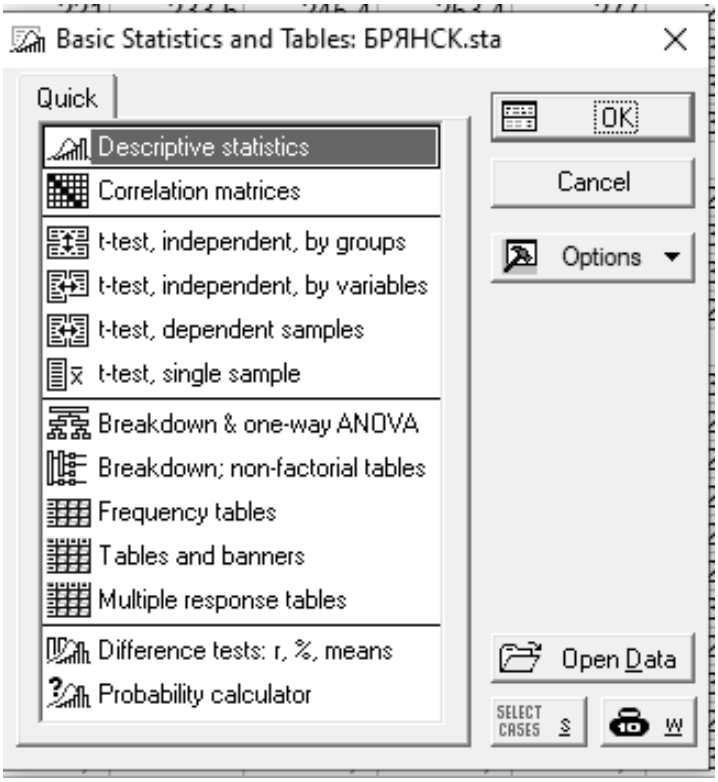

Рисунок 1 – Модуль проверки статистических гипотез

В таблице 2 представлены алгоритмы проверки параметрических гипотез (что означает, что выборки извлечены из нормально распространенных генеральных совокупностей).

Таблица 2 – Проверка гипотез

Гипотеза	H_0	Критерий	Примечание
Равенство дисперсии числовому параметру $X \in N(\mu;\sigma^2)$	$\sigma^2 = \sigma_0^2$	$X^2 = \frac{(n-1)S^2}{\sigma_0^2}$	–
Равенство математического ожидания числовому параметру $X \in N(\mu;\sigma^2)$	$\mu = \mu_0$	$U = \frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$	Дисперсия (генеральная) известна
		$T = \frac{\bar{x} - \mu_0}{\sqrt{\frac{S^2}{n-1}}}$	Дисперсия (генеральная) неизвестна

Окончание таблицы 2

Гипотеза	H_0	Критерий	Примечание
Равенство дисперсий $X \in N(\mu_x;\sigma^2)$ $Y \in N(\mu_y;\sigma^2)$	$D(X) = D(Y)$	$F = \frac{S_x^2}{S_y^2}$	В числителе всегда большая из двух дисперсий
Равенство математических ожиданий $X \in N(\mu_x;\sigma^2)$ $Y \in N(\mu_y;\sigma^2)$	$M(X) = M(Y)$	$U = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{D(X)}{n} + \frac{D(Y)}{m}}}$	Дисперсии (генеральные) известны из предшествующих наблюдений или определены теоретически
		$U = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$	Дисперсии (генеральные) неизвестны, но равны (предварительно проверить H_0: $D(X) = D(Y)$)
		$T = \frac{\bar{x} - \bar{y}}{\sqrt{(n-1)S_x^2 + (m-1)S_y^2}} \times \times \sqrt{\frac{nm(n+m-2)}{n+m}}$	Дисперсии (генеральные) неизвестны и не равны
Равенство вероятностей	$p_X = p_Y$	$Z = \frac{\frac{m_x}{n_x} + \frac{m_y}{n_y}}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}}$	$\bar{p} = \frac{m_x + m_y}{n_x + n_y}$

Задание для самостоятельной работы

На основании данных из открытых источников [7, 8] сформулировать и проверить гипотезы о статистически значимых различиях в выборках, распределение которых отлично от нормального. Приветствуется использование языков программирования R, Python. 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» [4].
2 Получить задание у преподавателя.
3 Реализовать задание.
4 Дать обоснование полученного решения.
5 Сделать выводы по результатам исследований.
6 Оформить отчет.

Требования к отчету

1 Цель работы.
2 Постановка задачи.
3 Результаты выполнения задания.
4 Выводы.

Задание

Анализируются данные, являющиеся случайной выборкой из записей о перепродажах домов, совершенных между 15 февраля и 30 апреля 1993 г. Информация предоставлена Советом риэлтеров (Albuquerque Board of Realtors) Альбукерка, США. Имеется 117 наблюдений.

Описание переменных:

– *PRICE* = продажная цена в сотнях долларов;
– *SQFT* = площадь в квадратных футах;
– *AGE* = возраст дома (количество лет);
– *FEATS* = количество дополнительных удобств из 11 возможных: dishwasher, refrigerator, microwave, disposer, washer, intercom, skylight(s), compactor, dryer, handicap fit, cable TV access;
– *NE* = дом расположен в престижном районе на северо-востоке города (1) или нет (0);
– *CUST* = тип постройки: был ли дом обычной постройки или нет;
– COR = как расположен дом: на углу (1) или нет (0).
– *TAX* = величина налогов за владение домом ($).

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

$$R = \begin{pmatrix} 1 & r_{12} & \dots & r_{1k} \\ r_{21} & 1 & \dots & r_{2k} \\ \dots & \dots & \dots & \dots \\ r_{k1} & \dots & \dots & 1 \end{pmatrix},$$

где r_{ij} – парные коэффициенты корреляции между i и j признаками, статистическая значимость которых определяется с помощью статистики Стьюдента

$$t_r = r_{ij} \cdot \sqrt{\frac{n-2}{1-r_{ij}^{\,2}}}\,.$$

Данные коэффициенты могут быть определены в модуле Basic Statistics and table [6]. 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 [6], статистическая значимость которых определяется аналогично (рисунок 2).

Задание для самостоятельной работы

На основании данных из открытых источников [7, 8] сформулировать и проверить гипотезы о наличии статистически значимой связи между количественными признаками. Приветствуется использование языков программирования R, Python. Результат исследования оформить в виде выступления на научной конференции.

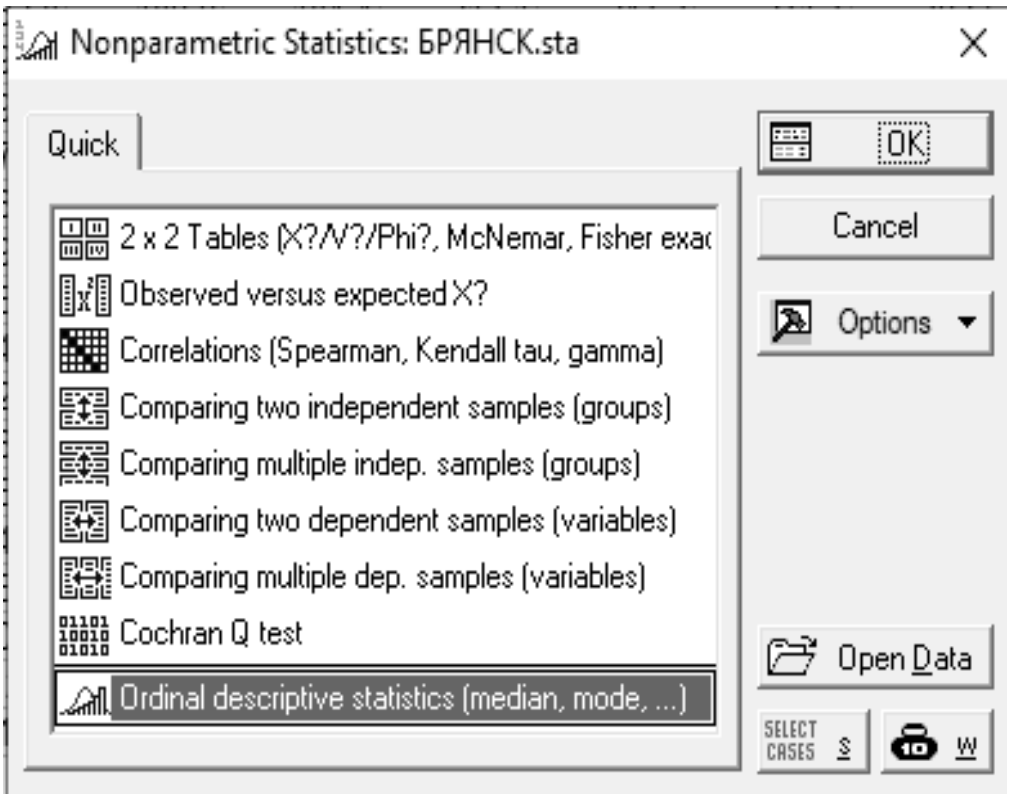

Рисунок 2 – Модуль Непараметрические статистики

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

Требования к отчету

1 Цель работы работы.
2 Постановка задачи.
3 Результаты выполнения задания.
4 Выводы.

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

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

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

α_i – неизвестные коэффициенты, подлежащие определению.

Метод наименьших квадратов (МНК) для оценки вектора параметров множественной линейной регрессионной модели предполагает применение следующей формулы:

$$\alpha = \left(X^T X \right)^{-1} \cdot X^T \cdot Y.$$

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. Ошибки ε_j распределены по нормальному закону с постоянной дисперсией ($\sigma^2 = \text{const}$).

Коэффициент множественной детерминации R показывает долю вариации результирующего признака, обусловленного вариацией факторных признаков.

Проверка значимости коэффициента детерминации осуществляется с помощью F-распределения:

$$F = \frac{R^2 / L}{(1 - R^2) / (n - L - 1)}; \quad \nu_1 = L; \quad \nu_2 = n - L - 1,$$

где L – количество фиксированных признаков.

Для определения значимости коэффициента α_j используется статистика t_j, имеющая распределение Стьюдента

$$t_j = \frac{\alpha_j}{S_{\alpha j}},$$

где $S_{\alpha j}$ – оценка ошибки j-го коэффициента,

Доверительный интервал для значимого коэффициента α_j

$$[\alpha_j - t_{\text{крит}} \cdot S_{\alpha_j}; \alpha_j + t_{крит} \cdot S_{\alpha_j}].$$

Доверительный интервал для результативного признака

$$[\hat{y}_p - t_{крит} \cdot S_{\hat{y}};\ \hat{y}_p + t_{крит} \cdot S_{\hat{y}}],$$

где $S_{\hat{y}} = S\sqrt{X_p^T \cdot C^{-1} \cdot X_p}$; $S^2 = \dfrac{\sum \varepsilon_j^2}{n-m-2}$; $C^{-1} = (X^T X)^{-1}$,

ошибка уравнения

$$S^2 = \frac{\sum \varepsilon_j^2}{n-m-2},$$

где X_p – матрица конкретных прогнозных значений независимых переменных;

ε_j^2 – квадрат отклонения эмпирического от теоретического значения результативного признака, $\varepsilon_j^2 = (Y_i - \hat{\hat{Y}}_i)^2$;

n – размерность X;

m – количество коэффициентов уравнения регрессии.

Доверительный интервал для прогнозного значения

$$[\hat{y}_p - t_{крит} \cdot S\sqrt{1 + X_p^T \cdot C^{-1} \cdot X_p};\ \hat{y}_p + t_{крит} \cdot S\sqrt{1 + X_p^T \cdot C^{-1} \cdot X_p}],$$

где $\hat{y}_p$ – теоретическое значение результативного признака, полученное с помощью уравнения регрессии.

Для обнаружения гетероскедастичности остатков используют тест ранговой корреляции Спирмена.

Все расчеты выполняются в модуле Multiply Regression программы Statistica (рисунок 3).

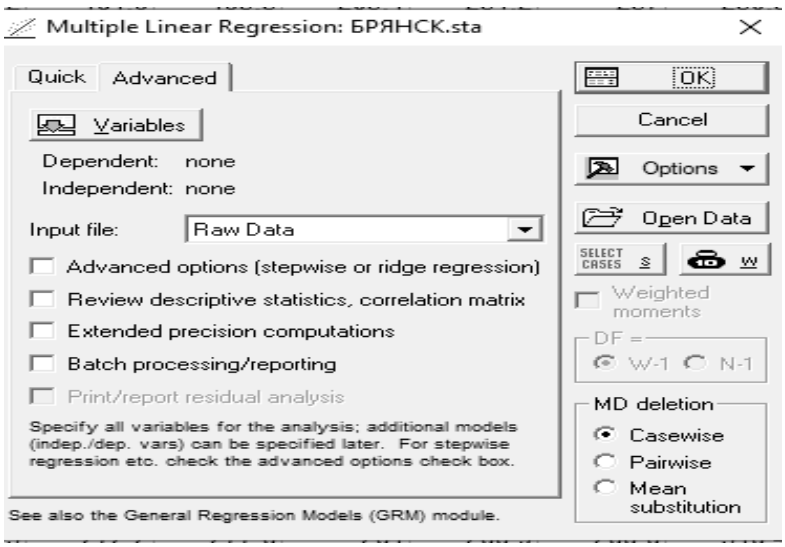

Рисунок 3 – Модуль Множественная регрессия

Задание для самостоятельной работы

На основании данных из открытых источников [7, 8] оценить модель линейной регрессии между количественными признаками. Приветствуется использование языков программирования R, Python. 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, продемонстрировать вариативность методов кластеризации.

Порядок выполнения работы

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

Требования к отчету

1 Цель работы.
2 Постановка задачи.
3 Результаты выполнения задания.
4 Выводы.

Задание

Имеются данные о приеме на работу на некоторое предприятие. 18 претендентов прошли 10 различных тестов в кадровом отделе. Максимальная оценка, которую можно было получить на каждом из тестов, составляет 10 баллов, минимальная – 1. Проверялись следующие качества:

1) память на числа;
2) умение решать математические задачи;
3) находчивость при прямом диалоге;
4) умение составлять алгоритмы;
5) уверенность во время выступления;
6) командный дух;
7) находчивость;
8) сотрудничество;
9) признание в коллективе;
10) сила убеждения.

Результаты теста хранятся в файле assess.dat в переменных *t*1 – *t*10 соответственно. В файле присутствуют также переменные с номером и фамилией участника.

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 [6] провести классификацию объектов. Сравнить решения, полученные методом иерархического кластерного анализа и методом *к*-средних.

Данные обсуждались в книге Бююль Цефель SPSS.

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.

Сравнить полученные результаты. Сделать выводы.

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 цен-

тров 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).

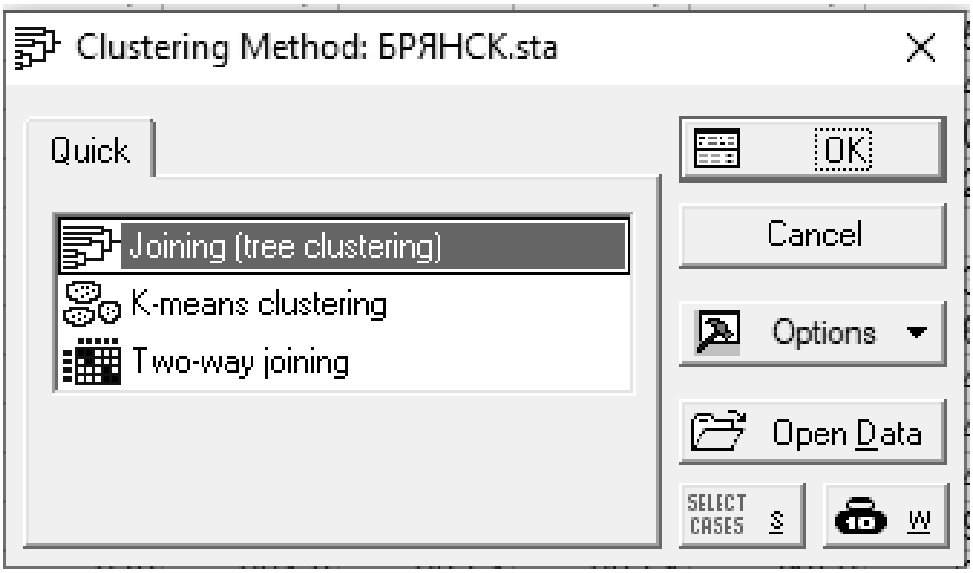

Рисунок 4 – Модуль Кластерный анализ

Чаще всего близость объектов x и y измеряется с помощью следующих метрик расстояния, если их характеристики измерены в интервальной шкале:

– расстояние Евклида: $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$;

– расстояние Манхэттена: $\sum_{i=1}^{n} |x_i - y_i|$;

– расстояние Чебышева: $\max_{i=1,N} |x_i - y_i|$.

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 [7, 8] выполнить разбиение на группы объектов, характеризующихся набором признаков, с помощью кластерного анализа. Приветствуется использование языков программирования R, Python. 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 «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.

Требования к отчету

1 Цель работы.
2 Постановка задачи.
3 Результаты выполнения задания.
4 Выводы.

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 – объект x можно отнести к классу $y = 1$, если предсказанная моделью вероятность $P\ (Y = 1/x) = 0{,}5$, и к классу $y = 0$ – в противном случае. Получающиеся при этом правила классификации являются линейными классификаторами.

Логистическая функция имеет следующий вид:

$$p = M(Y = 1/x_i) = \frac{1}{1 + e^{-Z}},$$

где Z – линейная комбинация классификаторов,

$$Z = F(x_1, x_2, \ldots, x_n).$$

Подбор коэффициентов регрессионного уравнения осуществляется на обучающей выборке с помощью метода максимального правдоподобия в модуле Nonlinear Estimation-Quick Logit regression в программе Statistica (рисунок 5).

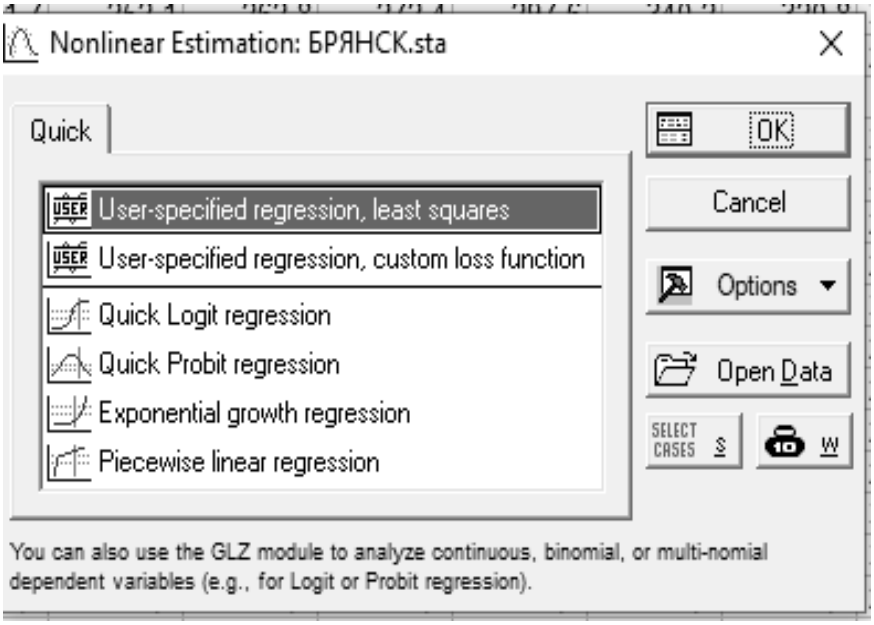

Рисунок 5 – Модуль Логистическая регрессия

Для оценки качества классификатора в программе может быть рассчитана четырехпольная классификационная таблица 3 (сопряженности).

Таблица 3 – Таблица сопряженности

Решение по тестируемому методу	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)

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

Специфичность (Specificity) – доля истинно отрицательных случаев

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

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 логистической регрессии служит ROC-кривая, а площадь под ней (AUC) является показателем качества классификационной модели (таблица 4).

Таблица 4 – 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 [7, 8] построить правило классификации объектов, характеризующихся набором признаков, с помощью логистической регрессионной модели. Оценить качество классификатора, построить ROC-кривую. Приветствуется использование языков программирования R, Python. 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? В каких диапазонах он может изменяться?

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.

Требования к отчету

1 Цель работы.
2 Постановка задачи.
3 Результаты выполнения задания.
4 Выводы.

Задание

1 Изучить теоретический материал по теме «Дискриминантный анализ» [4].

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) [6].

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.

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

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

b_i и 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.

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, то есть те переменные, чей вклад в 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 (рисунок 6) [6].

4 Интерпретация результатов.

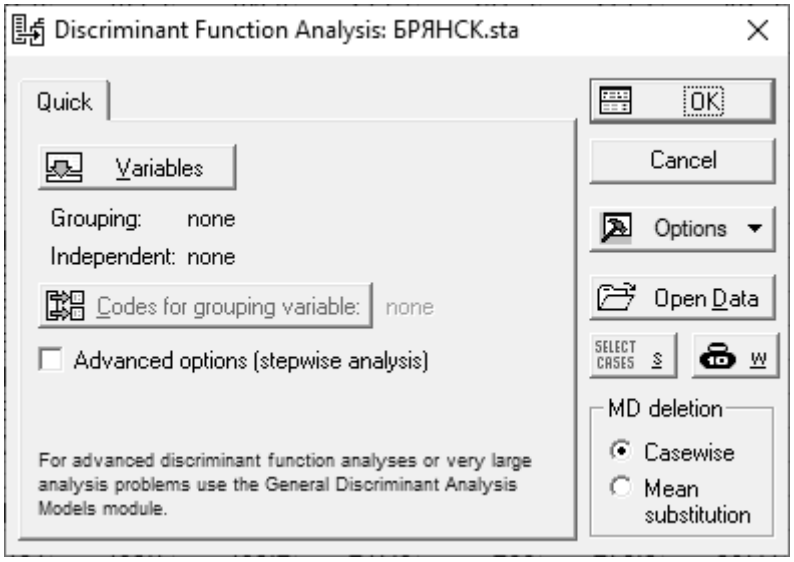

Рисунок 6 – Модуль Дискриминантный анализ

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).

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.

Требования к отчету

1 Цель работы.
2 Постановка задачи.
3 Результаты выполнения задания.
4 Выводы.

Задание

1 Изучить теоретический материал по темам «Факторный анализ как метод редукции данных», «Измерение и оценка факторов», «Применение факторного анализа к задачам классификации» [4, 6].

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. Для этого необходимо:

– 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:

$$X_i = \sum_{k=1}^{m} a_{ik} F_k + a_i U_i, \quad i = 1, 2, ..., n,$$

где U_i – некоторые «добавки», введение которых обусловлено строгим равенством, с одной стороны, и тем фактом, что m-мерный базис из факторов не обязательно окажется полным для исходного описания явления через n векторов-признаков. Факторы F_k называются общими факторами, а переменные U_i – специфическими факторами. Значения a_{ik} называются факторными нагрузками.

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

3 Как оценивается значимость модели факторного анализа?

4 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.

Требования к отчету

1 Цель работы.

2 Постановка задачи.

3 Результаты выполнения задания.

4 Выводы.

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.

Fertilit – показатель рождаемости.

Agriculture – процент мужчин в провинции, работающих в сельском хозяйстве.

Examination – процент призывников провинции, получивших высшие оценки на экзамене при поступлении в армию.

Education – процент призывников провинции, чье образование превышает уровень начальной (primary) школы.

Catholic – процент католиков.

Infant_Mortality – детская смертность, процент проживших меньше одного года.

Значения переменных *Examination* и *Education* являются средними значениями за 1887, 1888 и 1889 гг.

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. – Москва: Специалист, 2016. – 100 с.

2 [Электронный ресурс]. – Режим доступа: https://statanaliz.info/statistica/proverka-gipotez/kriterij-soglasiya-pirsona-khi-kvadrat/. – Дата доступа: 20.04.2020.

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

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

5 Центр справки и обучения Office [Электронный ресурс]. – Режим доступа: https://support.office.com. – Дата доступа: 27.12.2019.

6 StatSoft, Inc. (2012). Электронный учебник по статистике. Москва, StatSoft. – Режим доступа: http://www.statsoft.ru/home/textbook/default.htm. – Дата доступа: 27.03.2020.

7 [Электронный ресурс]. – Режим доступа: http://opendata.by/. – Дата доступа: 27.03.2020.

8 [Электронный ресурс]. – Режим доступа: https://www.kaggle.com/. – Дата доступа: 27.03.2020.