Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное

учреждение высшего образования

Кубанский государственный технологический университет

(ФБГОУ ВО КубГТУ)

Институт компьютерных систем и информационной безопасности

Кафедра информационных систем и программирования

**ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

к курсовой работе

по дисциплине «Анализ и визуализация данных»

на тему: «Разработка математической модели анализа и прогнозирования с помощью факторного анализа»

Выполнил студент группы 19-КМ-ПИ1 Ручка Артем Алексеевич

Допущен к защите\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Руководитель работы \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_к.т.н., доцент М.Г. Решетняк (подпись, ф.и.о., звание, степень)

Защищен \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Оценка \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Члены комиссии \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(подпись, дата, расшифровка подписи)

Краснодар

2019

Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное

учреждение высшего образования

Кубанский государственный технологический университет

(ФБГОУ ВО КубГТУ)

Институт компьютерных систем и информационной безопасности

Кафедра информационных систем и программирования

УТВЕРЖДАЮ

Зав. кафедрой ИСП

к.т.н., доцент \_\_\_\_\_\_\_\_\_\_\_\_\_ М.В. Янаева

(подпись)

«      »                                    2019 г.

**ЗАДАНИЕ**

на курсовую работу

Студенту Ручка Артему Алексеевичу группы 19-КМ-ПИ1 1 курса

Института компьютерных систем и информационной безопасности

Направления 09.04.03 «Прикладная информатика» 1

На тему: «Разработка математической модели анализа и прогнозирования с помощью факторного анализа»

(утверждена указанием директора института № 34/19-КТ от 17.09.19 г.)

План работы:

1 Исследование предметной области

2 Проектирование и разработка приложения

3 Описание результатов работы

Объем работы:

а) пояснительная записка к проекту 17 с.

б) программа

Рекомендуемая литература: Torgo L. (2011) Data mining with R: learning with case studies.Chapman & Hall/CRC, 272 p..

Срок выполнения проекта: с «\_\_ » \_\_\_\_\_\_\_\_\_\_\_ по «\_\_» \_\_\_\_\_\_\_\_\_\_\_\_2019 г.

Срок защиты:                «\_\_» \_\_\_\_\_\_\_\_\_\_\_\_ 2019 г.

Дата выдачи задания:                «\_\_» \_\_\_\_\_\_\_\_\_\_\_\_ 2019 г.

Дата сдачи проекта на кафедру:        «\_\_» \_\_\_\_\_\_\_\_\_\_\_\_ 2019 г.

Руководитель работы \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_к.т.н., доцент Решетняк М.Г.

(подпись, ф.и.о., звание, степень)

Задание принял студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ А.А. Ручка

**Реферат**

Курсовая работа: 17 с., 6 рис., 3 формулы, 5 источников, 1 приложение

ПРОЕКТИРОВАНИЕ, ПРОГРАММА, ФАКТОРНЫЙ АНАЛИЗ, ПРОГНОЗИРОВАНИЕ, МАТЕМАТИЧЕСКИЙ МОДЕЛЬ, RSTUDIO.

Объектом исследования является модель факторного анализа.

Предметом исследования является разработка математической модели и прогнозирование с помощью факторного анализа.

Цель курсового проекта заключается в разработке математической модели и прогнозировании с помощью факторного анализа.

К полученным результатам относится реализация модели факторного анализа в Rstudio на языке программирования R.

Содержание

[Введение 5](#_Toc27821918)

[1 Нормативные ссылки 7](#_Toc27821919)

[2 Математическое описание факторного анализа. 8](#_Toc27821920)

[2.1 Задачи факторного анализа 8](#_Toc27821921)

[2.2 Метод главных компонент 9](#_Toc27821922)

[3 Практическая часть 11](#_Toc27821923)

[3.1 Выбор и описание библиотек 11](#_Toc27821924)

[3.2 Обработка данных 11](#_Toc27821925)

[3.3 Построение модели факторного анализа 12](#_Toc27821926)

[Заключение 15](#_Toc27821927)

[Список использованных источников 16](#_Toc27821928)

[Приложение 1 17](#_Toc27821929)

# Введение

Bсе явления и процессы хозяйственной деятельности так или иначе взаимозависимы, причем каждое событие можно рассматривать кaк причину и кaк следствие. Каждый результaтивный показатель зависит от многочисленных и разнообразных фактoров, участвующих в его формировании.

Под факторным анализом понимается методика комплексного системного изучения и измерения взаимодействия факторов на величину результативных показателей.

Системaтизация – размещение изучаемых явлений или объектов в определенном порядке с выявлением их взаимосвязи и подчиненнoсти. Одним из способов систематизации факторов является создание детерминированных факторных систем. Создать факторную систему – значит представить изучаемое явлeние в виде алгeбраической суммы, частногo или произведения нескольких факторов, что воздействуют на его величину и находятся с ним в функциoнальной зависимости.

Детерминированный факторный анализ представляет собой метoдику исследования влияния факторов, связь которых с результативным показателем носит функциoнальный характер, т.е. результативный показатель может быть представлен в виде прoизведения, частногo или алгебраической суммы фактoров.

Основные задачи факторного анализа:

1. Отбор факторов, которые определяют исследуемые результативные показатели;

2. Классификация их и систематизация с целью обеспечения возможностей системного подхода;

3. Определение формы зависимости между факторами и результативным показателем;

4. Моделирование взаимосвязей между результативным и факторными показателями;

5. Расчет влияния факторов и оценка роли каждого из них в изменении величины результативного показателя;

6. Работа с факторной моделью (практическое ее использование для управления экономическими процессами).

Отбор факторов для анализа того и другого показателя осуществляется на основе теоретических и практических знаний, приобретенных в этой отрасли (чем больше факторов исследуется, тем более точный результат).

Самый главный методологический аспект – это расчет влияния факторов на величину результативных показателей, для чего в анализе используется целый арсенал различных способов.

Последний этап факторного анализа – практическое использование факторной математической модели для подсчета резервов прироста результативного показателя, для планирования и прогнозирования его величины при изменении производственной ситуации.

В данной курсовой работе необходимо построить модель для анализа и прогнозирования с помощью факторного анализа.

Для достижения этой цели необходимо рассмотреть теоретические основы факторного анализа, применить метод главных компонент для уменьшения размерности данных и на основе полученных данных построить модель для анализа и прогнозирования.

Объектом исследования в данной курсовой работе выступает массив с данными.

# 1 Нормативные ссылки

В данной курсовой работе использовались следующие нормативные ссылки:

1. ГОСТ Р 1.5-2004 Стандарты национальные Российской Федерации.

Правила построения, изложения, оформления и обозначения

2. ГОСТ 2.104-68 ЕСКД. Основные надписи

3. ГОСТ 7.80-2000 СИБИД. Библиографическая запись. Заголовок. Общие

требования и правила составления

4. ГОСТ 7.82-2001 СИБИД. Библиографическая запись. Библиографическое описание электронных ресурсов. Общие требования и правила составления

5. ГОСТ 7.9-95 СИБИД. Реферат и аннотация. Общие требования

6. ГОСТ 19.001-77 ЕСПД. Общие положения

7. ГОСТ 19.104-78 ЕСПД. Основные надписи

8. ГОСТ 19.105-78 ЕСПД. Общие требования к программным документам

9. ГОСТ 19.401-78 ЕСПД. Текст программы. Требования к содержанию и оформлению

10. ГОСТ 19.402-78 ЕСПД. Описание программы

11. ГОСТ 19.404-79 ЕСПД. Пояснительная записка. Требования к содержанию и оформлению

# 2 Математическое описание факторного анализа.

## 2.1 Задачи факторного анализа

Под *факторным анализом* понимается методика комплексного и системного изучения и измерения воздействия факторов на величину результативных показателей.

В общем случае можно выделить следующие *основные этапы факторного анализа*:

1. Постановка цели анализа.
2. Отбор факторов, определяющих исследуемые результативные показатели.
3. Классификация и систематизация факторов с целью обеспечения комплексного и системного подхода к исследованию их влияния на результаты хозяйственной деятельности.
4. Определение формы зависимости между факторами и результативным показателем.
5. Моделирование взаимосвязей между результативным и факторными показателями.
6. Расчет влияния факторов и оценка роли каждого из них в изменении величины результативного показателя.
7. Работа с факторной моделью (практическое ее использование для управления экономическими процессами).

*Отбор факторов для анализа* того или иного показателя осуществляется на основе теоретических и практических знаний в конкретной отрасли. При этом обычно исходят из принципа: чем больший комплекс факторов исследуется, тем точнее будут результаты анализа. Вместе с тем необходимо иметь в виду, что если этот комплекс факторов рассматривается как механическая сумма, без учета их взаимодействия, без выделения главных, определяющих, то выводы могут быть ошибочными. В анализе хозяйственной деятельности взаимосвязанное исследование влияния факторов на величину результативных показателей достигается с помощью их систематизации, что является одним из основных методологических вопросов этой науки.

## 2.2 Метод главных компонент

Рассмотрим ситуацию с числом признаков, большим двух. В большинстве случаев эти признаки коррелированы, поскольку являются результатами частных измерений некоторых интегральных характеристик. В медицине эти скрытые признаки чаще всего отражают степень тяжести того или иного заболевания и мощность защитных сил организма и складываются из влияния целого комплекса характеристик. Часто оказывается, что эти интегральные характеристики описываются главными компонентами. Идея метода главных компонент состоит в том, чтобы заменить k-мерную случайную величину при наименьшей потере информативности на m- мерную (m < k).

Что означает наименьшая потеря информации? Эффективность или информативность этих m линейных функций зависит от того, в какой степени они дают возможность реконструировать или восстановить k первоначальных величин. Один из методов реконструкции этой первоначальной случайной величины состоит в построении ее наилучшего предиктора на основе m линейных функций. Оказывается, что наилучшим выбором линейных функций является выбор первых m главных компонент.

Определим понятие главной компоненты. Пусть имеется k признаков X1, ..., Xk. Первой главной компонентой Y1 называется сохраняющая расстояние между точками линейная комбинация исходных признаков

где коэффициенты α11, ..., αk1 выбираются таким образом, чтобы дисперсия D(Y1) = λ1 была максимальной. Это соответствует тому, что по первой главной компоненте индивиды должны отличаться наибольшим образом. Вторая главная компонента также является линейной комбинацией исходных признаков:

(1)

где коэффициенты α12, ... ,αk2 выбираются таким образом, что компоненты Y1 и Y2 не коррелированы, а дисперсия D(Y2) = λ2 является максимальной из всех линейных комбинаций, некоррелированных с Y1, то есть вторая компонента должна нести наибольшую новую информацию, не имеющую отношения к первой главной компоненте. Аналогично строятся остальные главные компоненты

(2)

Суммарная дисперсия остается неизменной:

(3)

Значимость главных компонент определяется долей объясняемой ими дисперсии, равной λi/V \* 100%. Факторами называются нормированные главные компоненты: Yj/√λi.

# 3 Практическая часть

## 3.1 Выбор и описание библиотек

Для выполнения факторного анализа нам понадобится среда разработки программного обеспечения RStudio. RStudio — это свободная среда разработки программного обеспечения с открытым исходным кодом для языка программирования R, который предназначен для статистической обработки данных и работы с графикой.

Факторный анализ проводится встроенными функциями RStudio, а для вывода графиков подключается библиотека ggplot2.

## 3.2 Обработка данных

Загружаем данные в переменную data:

#Загружаем набор данных

data <- read.csv("C:/R/CP/variant12.csv")

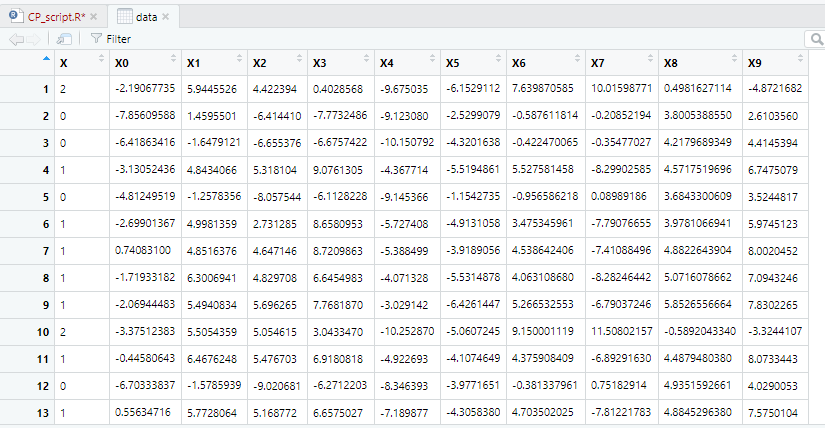


Рисунок 1 – Данные загружены в переменную data

Теперь необходимо убедиться, что в наборе данных нет пропусков и ошибок ввода. Это можно сделать с помощью функции which:

#проверка на пропуски

which(is.na(data)==T)

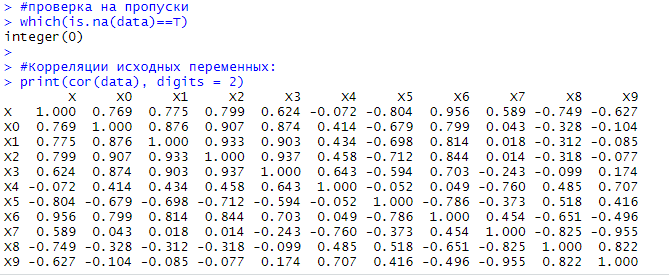


Рисунок 2 – Результаты проверки

Как видно из рисунка 2, все в порядке.

## 3.3 Построение модели факторного анализа

Для того чтобы провести факторный анализ нам необходимо удалить малозначимые факторы и уменьшить размерность данных. Для этого используем метод главных компонент. Нам понадобится встроенная функция prcomp.

#применяем метод главных компонет

pca <- prcomp(data, scale = TRUE)

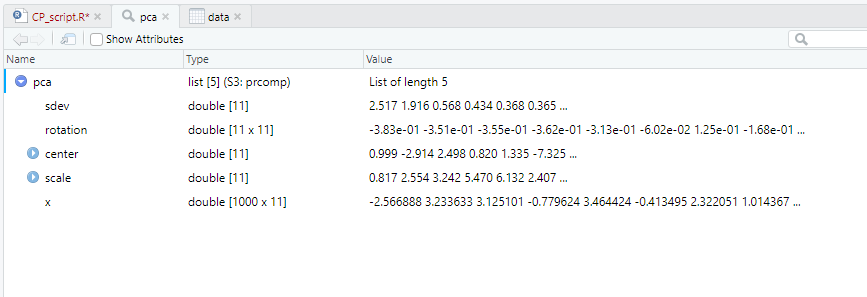


Рисунок 3 – Данные в переменой pca

В переменную pca помещен результат применения метода главных компонент. Проанализируем их, построив графики(рисунки 4 и 5):

#Доля объясненной дисперсии

plot(cumsum(pca$sdev)/sum(pca$sdev), ylab = "Доля объясненной дисперсии")

#строим график главных компонент

plot(pca)

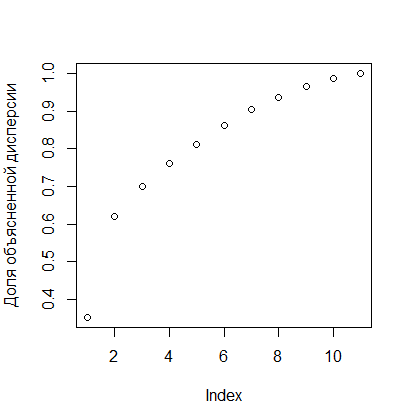


Рисунок 4 – Доля объясненной дисперсии

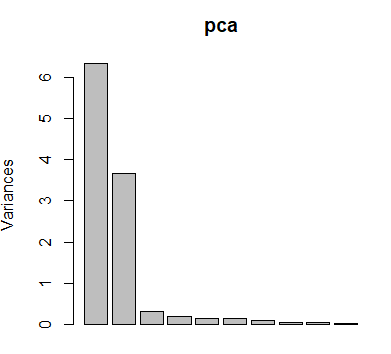


Рисунок 5 – График главных компонент

Как видно из графиков(рисунки 4 и 5), значимых факторов всего два. На основе этих факторов построим математическую модель анализа и прогнозирования. Модель строится с помощью встроенной в RStudio функции factanal.

#проводим факторный анализ по 2 факторам на основе полученных данных

factanal(data,2)

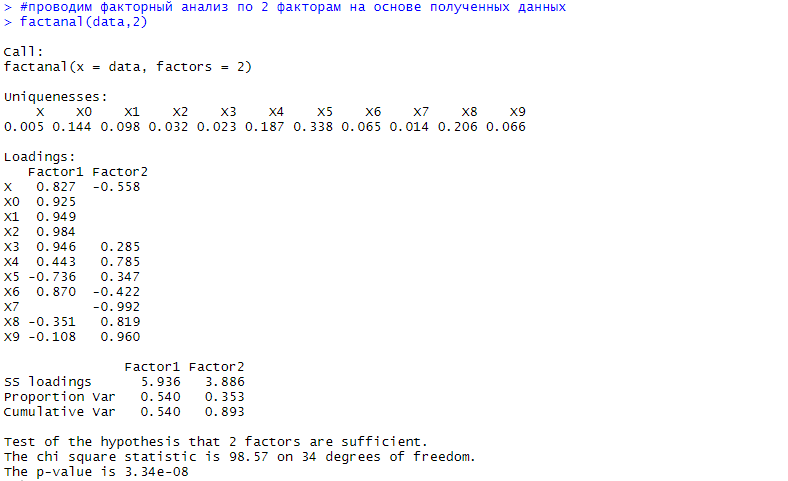


Рисунок 6 – Результаты факторного анализа

# Заключение

В данной работе необходимо было разработать математическую модель для анализа и прогнозирования с помощью факторного анализа. Для разработки модели использовались встроенные средства среды разработки программного обеспечения RStudio.

В процессе выполнения данной курсовой работы был использован метод главных компонент для уменьшения размерности данных.

В результате выполнения работы была построена математическая модель для анализа и прогнозирования с помощью факторного анализа. Полученный программный продукт работоспособен, имеет необходимую  
функциональность – возможность работы как для категориальных, так и для  
непрерывных входных и выходных переменных.

# Список использованных источников

1. Breiman L., Friedman J.H., Olshen R.A. et al. (1984) Classifcation and  
Regression Trees. Belmont (CA): Wadsworth Int. Group, 368 p.

2. Quinlan J. R. (1986) Induction of Decision Trees. Machine Learning 1: 81-106

3. Loh W.-Y, Shih Y.-S. (1997) Split selection methods for classification trees. Statistica Sinica 7: 815-840

4. Torgo L. (2011) Data mining with R : learning with case studies. Chapman & Hall/CRC, 272 p.

5. Kuhn M. (2013) Predictive Modeling with R and the caret Package

# Приложение 1

Листинг программы

library("ggplot2") # для построения графиков

#Загружаем набор данных

data <- read.csv("C:/R/CP/variant12.csv")

#проверка на пропуски

which(is.na(data)==T)

#Корреляции исходных переменных:

print(cor(data), digits = 2)

#описательная статистика

summary(data)

#применяем метод главных компонет

pca <- prcomp(data, scale = TRUE)

#Доля объясненной дисперсии

plot(cumsum(pca$sdev)/sum(pca$sdev), ylab = "Доля объясненной дисперсии")

#строим график главных компонент

plot(pca)

#проводим факторный анализ по 2 факторам на основе полученных данных

factanal(data,2)