МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ

УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«Национальный исследовательский ядерный университет «МИФИ»

(НИЯУ МИФИ)

Институт Финансовых Технологий и Экономической Безопасности

Кафедра Финансового Мониторинга

Лабораторная работа №3:

По курсу «Макростатический анализ и прогнозирование»

Работу выполнил: студент группы С18-712: Луканов А.В.

Проверила: Домашова Д. В.

Москва 2021

# 1. Постановка задачи

Субъект РФ характеризуются следующими признаками:

X1 - соотношение мужчин и женщин (оценка на конец года; на 1000 мужчин приходится женщин);

Х2 – численность населения с денежными доходами ниже величины прожиточного минимума в процентах от общей численности населения субъекта Российской Федерации;

Х3 – общие коэффициенты брачности на 1000 человек населения

Х4 – среднедушевые денежные доходы населения (в месяц; рублей);

Х5 – коэффициенты миграционного прироста на 10 000 человек населения;

Х6 – зарегистрировано преступлений, связанных с незаконным оборотом наркотиков (на 1000 человек);

Х7 – зарегистрировано преступлений средней тяжести (на 1000 человек);

Х8 – зарегистрировано преступлений особой тяжести (на 1000 человек);

Х9 – предварительно расследовано преступлений, совершенных в состоянии алкогольного опьянения (на 1000 человек);

Х10 - зарегистрировано преступлений экономической направленности (на 1000 человек).

На основе предварительного экспертного анализа было выделено 6 групп субъектов. К первой группе отнесено 2 субъекта, ко второй – 8, к третьей – 4, к четвертой - 9, к пятой – 3, к шестой 3.

Ставится задача на основании статистических данных по показателям, соответствующим нужному варианту, снизить размерность признакового пространства методом главных компонент, обеспечив уровень информативности новой системы признаков не ниже 70%.

# 2. Ход работы.

Первым делом стандартизируем исходные данные, чтобы привести их к одному масштабу.

Если исходные признаки, по которым производится классификация объектов, имеют разные единицы измерения, то необходимо перейти к стандартизованным переменным одним из следующих способов:

; ; ; ; ; ,

где  - исходное значение j-го признака на *i*-ом объекте наблюдения;

 - нормированное значение исходного *j*-го признака на *i*-ом объекте наблюдения;

 - среднее значение *j*-го признака;

 - выборочное среднеквадратическое отклонение *j*-го признака;

 - максимальное значение *j*-го признака;

 - минимальное значение *j*-го признака.

Была рассчитана матрица корреляций для исходных данных. Матрицы представлены на Рисунках 1.1 и 1.2.

Изображение выглядит как текст, кроссворд, табло

Автоматически созданное описание

Рисунок 1.1 – Результат расчета корреляционной матрицы (Statistica).

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 1.2 – Результат расчета корреляционной матрицы (Python).

Теперь проверим гипотезу о незначимости матрицы корреляций. Для этого посчитаем собственные числа для данной матрицы. Результаты представлены на Рисунках 2.1, 2.2.

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 2.1 – Результаты расчета оценок собственных чисел корреляционной матрицы (Statistica)

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 3.1 – Результаты расчета оценок собственных чисел корреляционной матрицы (Python)

Затем были посчитаны наблюдаемое и критическое значение статистики Хи-квадрат, в результате гипотеза о незначимости корреляционной матрицы была отвергнута. Результаты представлены на Рисунках 3.1 и 3.2.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Hi2набл | Hi2крит1 | Hi2крит2 | alpha | k | n | det(R) = |
| 309,6353 | 82,86705 | 140,9166 | 0,05 | 11 | 85 | 0,021356 |

Рисунок 3.1– Проверка гипотезы о незначимости корреляционной матрицы в Excel



Рисунок 3.2– Проверка гипотезы о незначимости корреляционной матрицы в Python

Дальше были посчитаны доверительные интервалы для собственных значений корреляционной матрицы, результаты представлены на рисунках 4.1 и 4.2.

|  |  |  |
| --- | --- | --- |
| Предел 1 | Собственные числа | Предел 2 |
| 2,332621 | 3,038074 | 4,355218 |
| 2,119357 | 2,760312 | 3,957033 |
| 1,040047 | 1,354587 | 1,941863 |
| 0,862791 | 1,123724 | 1,61091 |
| 0,673148 | 0,876727 | 1,256829 |
| 0,504738 | 0,657385 | 0,942392 |
| 0,303441 | 0,395211 | 0,566553 |
| 0,24849 | 0,323640 | 0,463953 |
| 0,176508 | 0,229889 | 0,329557 |
| 0,125514 | 0,163472 | 0,234345 |

Рисунок 4.1 – Доверительные интервалы собственных чисел, посчитанные в Excel

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 4.2 – Доверительные интервалы собственных чисел, посчитанные в Python

Посчитаем информативность (Рисунок 5.1 и 5.2).

|  |  |
| --- | --- |
| Оценка информативности | |
| I\_1(z(x)) = | 0,276189 |
| I\_2(z(x)) = | 0,527126 |
| I\_3(z(x)) = | 0,65027 |
| I\_4(z(x)) = | 0,752427 |

Рисунок 5.1 – Оценка информативности (Excel)

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 5.2 – Оценка информативности (Python)

Информативность превышает 70% при использовании 4 главных компонент. Проверим гипотезу о достаточности 4 главных компонент с помощью критерия Кайзера (Рисунок 6.1 и 6.2).



Рисунок 6.1 – Критерий Кайзера (Excel)



Рисунок 6.2 – Критерий Кайзера (Python)

Построим график «осыпи». Результаты представлены на рисунках 7.1 и 7.2.

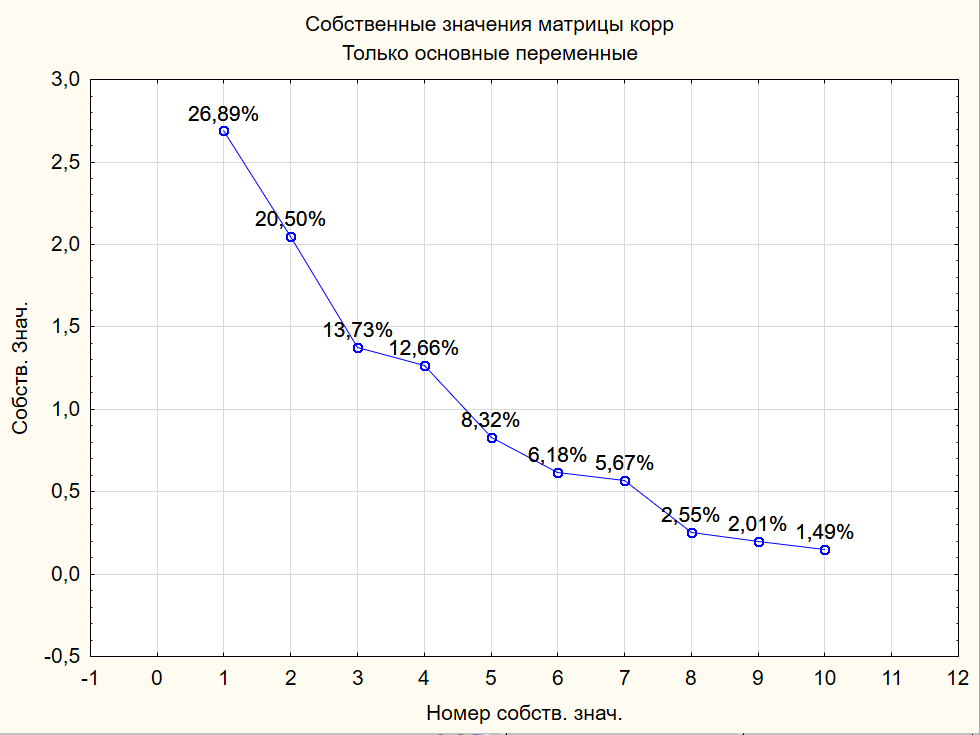


Рисунок 7.2 – График «осыпи» (Statistica)

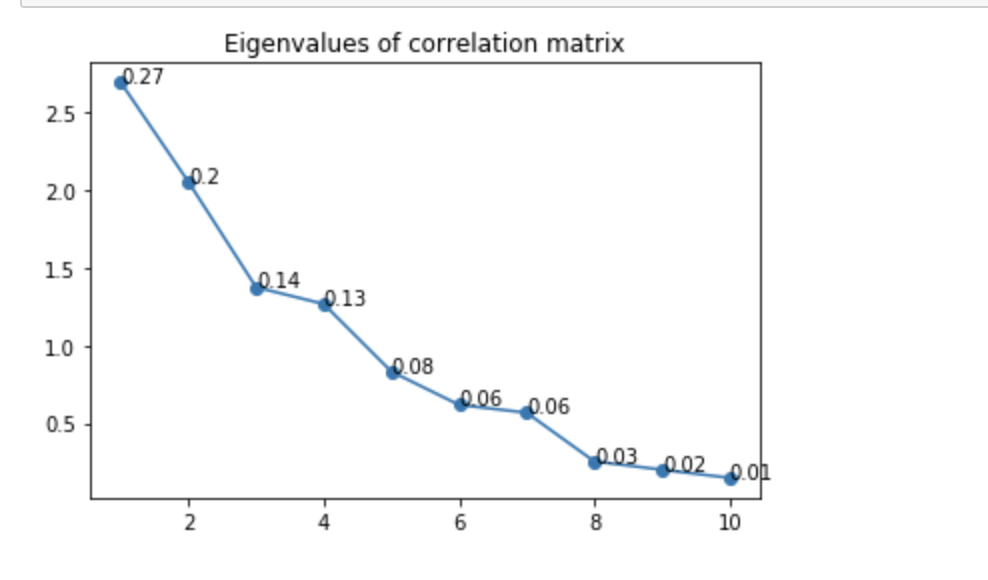


Рисунок 7.2 – График «осыпи» (Python)

Определим вклад каждой главной компоненты в суммарную дисперсию исходных признаков (Рисунок 8.1 и 8.2).

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 8.1 – Вклады главных компонент в суммарную дисперсию исходных признаков (Statictica)

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 8.2 – Вклады главных компонент в суммарную дисперсию исходных признаков (Python)

Рассчитаем собственные вектора корреляционной матрицы. Результаты представлены на рисунке 9.

Изображение выглядит как текст, кроссворд, табло

Автоматически созданное описание

Рисунок 9 – Результаты расчета собственных векторов корреляционной матрицы (Statistica)

После этого посчитаем коэффициенты линейного преобразования центрировано-нормированных исходных признаков (Рисунок 10).

Изображение выглядит как текст, кроссворд

Автоматически созданное описание

Рисунок 10.1 – Коэффициенты линейного преобразования центрировано-нормированных исходных признаков (Statistica)

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 10.2 – Коэффициенты линейного преобразования центрировано-нормированных исходных признаков (Python)

Рассчитаем матрицу нагрузок . Результаты представлены на рисунках 11.1 и 11.2.

Изображение выглядит как текст, кроссворд, окно

Автоматически созданное описание

Рисунок 11.1 – Результаты расчета элементов матрицы нагрузок (Statistica)

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 11.2 – Результаты расчета элементов матрицы нагрузок (Python)

Проведем анализ матрицы нагрузок. Так как расчеты проводятся на основании корреляционной матрицы, то элементы матрицы нагрузок являются коэффициентами корреляции исходных признаков и главных компонент. Как видно из таблицы, между исходными признаками и последними 6 главными компонентами не наблюдается тесной связи (не имеется значений >0,7) Это подтверждает правильность выделения только 4 первых главных компонент.

Первая главная компонента тесно отрицательно связана (модуль коэффициента корреляции больше 0,7) с четырьмя исходными признаками: x6 (Число преступлений, связанных с оборотом наркотиков), x7 (Число преступлений средней тяжести), x8 (Число преступлений особой тяжести), x9 (Число преступлений, совершенных в состоянии алкогольного опьянения). Поэтому первую главную компоненту можно интерпретировать как «Показатель преступности».

Вторая главная компонента тесно отрицательно связана с признаком x2 (Доля населения с доходами ниже прожиточного минимума) и отрицательно с признаком х4 (Среднедушевые доходы населения), поэтому вторую главную компоненту можно интерпретировать «Показатель бедности».

Третья главная компонента тесно положительно с признаком х10 (Число экономических преступлений), поэтому третью главную компоненту можно интерпретировать «Показатель экономической преступности»

Четвертая главная компонента тесно положительно с признаком х1 (Число женщин на 1000 мужчин), поэтому четвертую главную компоненту можно интерпретировать «Показатель соотношения полов»

Рассчитаем матрицу индивидуальных значений центрировано - нормированных главных компонент. Результаты представлены на рисунках 12.1 и 12.2.

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 12.1 - Фрагмент матрицы индивидуальных значений центрировано- нормированных главных компонент (Statistica)

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 12.2 - Фрагмент матрицы индивидуальных значений центрировано- нормированных главных компонент (Python)