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»**

|  |  |
| --- | --- |
| **Выполнил студент группы С21-703:** | Монастырский М. О. |
| **Проверил:** | Домашова Д.В. |

Москва, 2024

Оглавление

[1. Постановка задачи 3](#_Toc181650853)

[2. Выполнение работы 4](#_Toc181650854)

[3. Метод Уорда (МГК) 14](#_Toc181650855)

[4. Метод К-средних (МГК) 16](#_Toc181650856)

1. Постановка задачи

Субъекты РФ характеризуются социально-экономическими показателями, обозначение и наименование которых приведены в таблице 1. Значения показателей для 85 субъектов приведены в таблице А. 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%.

|  |  |
| --- | --- |
|  | Число дорожно-транспортных происшествий и пострадавших в них на 100 000 человек населения |
|  | Смертность населения старше трудоспособного возраста, на 100 000 человек населения соответствующего возраста |
|  | Продажа сильно алкогольной продукции населению(тысяч декалитров)/**на тыс населения** |
|  | Средняя Стоимость минимального (условного) набора потребительских товаров и услуг |
|  | Число спортивных сооружений/ **на тыс населения** |
|  | Доходы консолидированных бюджетов субъектов Российской Федерации / **на тыс населения** |
|  | Предварительно расследовано преступлений, совершенных в состоянии алкогольного опьянения/ **на тыс населения** |
|  | Среднедушевые доходы населения (в месяц), руб. |
|  | Численность студентов, обучающихся по программам бакалавриата, специалитета, магистратуры на 10 000 человек населения, всего |

Поскольку исходные признаки отличаются масштабом измерения, то будем рассматривать вектор центрировано-нормированных признаков и на основе исходной матрицы данных X рассчитаем оценку корреляционной матрицы .

2. Выполнение работы

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | Correlations (Переделка) Marked correlations are significant at p < ,05000 N=85 (Casewise deletion of missing data) | | | | | | | | | | |
| |  | | --- | | Means | | |  | | --- | | Std.Dev. | | |  | | --- | | X1 | | |  | | --- | | X2 | | |  | | --- | | X3 | | |  | | --- | | X4 | | |  | | --- | | X5 | | |  | | --- | | X6 | | |  | | --- | | X7 | | |  | | --- | | X8 | | |  | | --- | | X9 | |
| |  | | --- | | X1 | | 0,000000 | 1,000000 | 1,000000 | 0,424649 | 0,231616 | -0,162421 | 0,255123 | -0,179811 | 0,325123 | -0,275671 | -0,025242 |
| |  | | --- | | X2 | | 0,000000 | 1,000000 | 0,424649 | 1,000000 | 0,594993 | 0,021442 | 0,381429 | 0,196778 | 0,612704 | 0,061868 | -0,236869 |
| |  | | --- | | X3 | | 0,000000 | 1,000000 | 0,231616 | 0,594993 | 1,000000 | 0,004660 | 0,140411 | 0,371633 | 0,472637 | 0,483304 | -0,343512 |
| |  | | --- | | X4 | | -0,000000 | 1,000000 | -0,162421 | 0,021442 | 0,004660 | 1,000000 | -0,127318 | -0,131076 | 0,058640 | -0,107069 | -0,057333 |
| |  | | --- | | X5 | | -0,000000 | 1,000000 | 0,255123 | 0,381429 | 0,140411 | -0,127318 | 1,000000 | 0,165129 | 0,289761 | -0,011878 | -0,018994 |
| |  | | --- | | X6 | | -0,000000 | 1,000000 | -0,179811 | 0,196778 | 0,371633 | -0,131076 | 0,165129 | 1,000000 | 0,452447 | 0,750230 | -0,322548 |
| |  | | --- | | X7 | | 0,000000 | 1,000000 | 0,325123 | 0,612704 | 0,472637 | 0,058640 | 0,289761 | 0,452447 | 1,000000 | 0,138288 | -0,360360 |
| |  | | --- | | X8 | | 0,000000 | 1,000000 | -0,275671 | 0,061868 | 0,483304 | -0,107069 | -0,011878 | 0,750230 | 0,138288 | 1,000000 | -0,185402 |
| |  | | --- | | X9 | | -0,000000 | 1,000000 | -0,025242 | -0,236869 | -0,343512 | -0,057333 | -0,018994 | -0,322548 | -0,360360 | -0,185402 | 1,000000 |

Далее согласно алгоритму, предполагая, что выборка извлечена из нормально распределенной генеральной совокупности, на уровне значимости α = 0,05 проверим гипотезу о незначимости корреляционной матрицы.

H0: Rx = Е;

H1: Rx ≠ E.

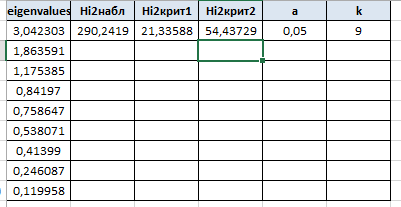
Для проверки гипотезы потребуются оценки собственных чисел корреляционной матрицы.

Расчет наблюдаемого значения проводится в Excel на основании полученных оценок собственных значений, которые приводятся на рисунке 2.

No. of active vars:9 No. of supplementary vars:0

No. of active cases: 85 No. of supplementary cases: 0

Eigenvalues: 3,04230 1,86359 1,17539 ,841970 ,758647 ...



Наблюдаемое значение рассчитывается по формуле:

χ2 = -(n - (2k+5))ln||,

где || - определитель матрицы , равный произведению оценок собственных чисел матрицы;

k – число факторов;

n – объем выборки.

Наблюдаемое значение составило χ 2набл =365,21. Критические значения χкр1 и 2 χкр2 определяются из уравнений:

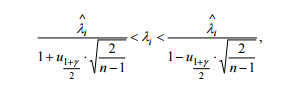


Для решения этих уравнений необходимо воспользоваться функцией ХИ2ОБР(вероятность, ) пакета Excel. Вероятность рассчитывается как 100\*(1-)% для χкр1 и 100\*()% для χкр2 . Число степеней свободы ν = . Критические точки принимают следующие значения:

Так как , то гипотеза  отвергается, матрица парных коэффициентов корреляции значима.

С вероятностью γ = 0,95 построим доверительные интервалы для собственных чисел.

Доверительный интервал для i-ого собственного числа λi при большом объеме выборки имеет вид:

,

где – квантиль уровня стандартного нормального распределения;

n – объем выборки.

Квантиль уровня q можно найти с помощью функции НОРМСТОБР(q) пакета Excel. Доверительные интервалы для собственных чисел имеют вид, представленный на рисунке 5.



Количество главных компонент можно посчитать с помощью Критерия Кайзера (отбираются факторы с собственными числами > 1, в данном примере их получается 3, то есть гипотеза о достаточности двух главных компонент не принимается).

Так как собственные числа корреляционной матрицы являются дисперсиями главных компонент, то оценка уровня информативности первых двух главных компонент составляет I2(z(x)) = что меньше требуемого уровеня 55%.

I3(z(x)) = что больше требуемого уровеня 55%.

Так же можно использовать Критерий каменистой осыпи, для этого в Statistica строится графическое представление собственных чисел («График осыпи» в русской Statistica).



Следует найти такое место на графике, где убывание собственных значений слева направо максимально замедляется. Предполагается, что справа от этой точки находится только "факториальная осыпь". В соответствии с этим критерием можно оставить в этом примере 3 главные компоненты

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Value number | Eigenvalues of correlation matrix, and related statistics (Переделка) Active variables only | | | |
| |  | | --- | | Eigenvalue | | |  | | --- | | % Total variance | | |  | | --- | | Cumulative Eigenvalue | | |  | | --- | | Cumulative % | |
| |  | | --- | | 1 | | 3,042303 | 33,80337 | 3,042303 | 33,8034 |
| |  | | --- | | 2 | | 1,863591 | 20,70656 | 4,905894 | 54,5099 |
| |  | | --- | | 3 | | 1,175385 | 13,05983 | 6,081279 | 67,5698 |
| |  | | --- | | 4 | | 0,841970 | 9,35522 | 6,923248 | 76,9250 |
| |  | | --- | | 5 | | 0,758647 | 8,42941 | 7,681895 | 85,3544 |
| |  | | --- | | 6 | | 0,538071 | 5,97856 | 8,219966 | 91,3330 |
| |  | | --- | | 7 | | 0,413990 | 4,59988 | 8,633955 | 95,9328 |
| |  | | --- | | 8 | | 0,246087 | 2,73430 | 8,880042 | 98,6671 |
| |  | | --- | | 9 | | 0,119958 | 1,33286 | 9,000000 | 100,0000 |

В первом столбце таблицы приведены оценки собственных чисел, в третьем столбце – накопленные значения собственных чисел, во втором и в четвертом столбцах – относительный вклад каждой главной компоненты в суммарную дисперсию и накопленный относительный вклад соответственно.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | Factor score coefficients, based on correlations (Переделка) | | | | | | | | |
| |  | | --- | | Factor 1 | | |  | | --- | | Factor 2 | | |  | | --- | | Factor 3 | | |  | | --- | | Factor 4 | | |  | | --- | | Factor 5 | | |  | | --- | | Factor 6 | | |  | | --- | | Factor 7 | | |  | | --- | | Factor 8 | | |  | | --- | | Factor 9 | |
| |  | | --- | | X1 | | -0,099603 | -0,404622 | 0,125102 | 0,355039 | 0,165920 | 0,218535 | 1,055717 | 0,36141 | 0,02097 |
| |  | | --- | | X2 | | -0,244564 | -0,236411 | -0,081429 | -0,071245 | 0,235784 | -0,134032 | -0,787373 | 1,25531 | -0,21471 |
| |  | | --- | | X3 | | -0,260293 | 0,019906 | -0,098344 | 0,218545 | 0,526677 | -0,553197 | -0,065727 | -1,00866 | 0,97611 |
| |  | | --- | | X4 | | 0,023939 | 0,001164 | -0,733590 | -0,507493 | 0,183499 | -0,008839 | 0,523574 | 0,18619 | 0,10259 |
| |  | | --- | | X5 | | -0,132555 | -0,203902 | 0,332394 | -0,717006 | -0,361299 | -0,539179 | 0,247968 | -0,31104 | -0,11645 |
| |  | | --- | | X6 | | -0,220847 | 0,312880 | 0,157550 | -0,174155 | -0,153242 | 0,476790 | 0,248657 | 0,53955 | 1,77302 |
| |  | | --- | | X7 | | -0,257981 | -0,116690 | -0,138922 | -0,088065 | -0,177286 | 0,860425 | -0,265795 | -0,89825 | -0,88019 |
| |  | | --- | | X8 | | -0,170054 | 0,396346 | 0,156388 | -0,036706 | 0,326578 | -0,136977 | 0,434359 | 0,28187 | -1,82772 |
| |  | | --- | | X9 | | 0,169775 | -0,080099 | 0,318770 | -0,444659 | 0,794889 | 0,457743 | -0,141215 | -0,17596 | 0,23236 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | Eigenvectors of correlation matrix (Переделка) Active variables only | | | | | | | | |
| |  | | --- | | Factor 1 | | |  | | --- | | Factor 2 | | |  | | --- | | Factor 3 | | |  | | --- | | Factor 4 | | |  | | --- | | Factor 5 | | |  | | --- | | Factor 6 | | |  | | --- | | Factor 7 | | |  | | --- | | Factor 8 | | |  | | --- | | Factor 9 | |
| |  | | --- | | X1 | | -0,173729 | -0,552363 | 0,135629 | 0,325780 | 0,144517 | 0,160303 | 0,679270 | 0,179285 | 0,007264 |
| |  | | --- | | X2 | | -0,426574 | -0,322732 | -0,088281 | -0,065374 | 0,205369 | -0,098317 | -0,506612 | 0,622722 | -0,074363 |
| |  | | --- | | X3 | | -0,454008 | 0,027175 | -0,106620 | 0,200534 | 0,458737 | -0,405788 | -0,042290 | -0,500369 | 0,338075 |
| |  | | --- | | X4 | | 0,041755 | 0,001589 | -0,795322 | -0,465670 | 0,159828 | -0,006484 | 0,336878 | 0,092364 | 0,035531 |
| |  | | --- | | X5 | | -0,231205 | -0,278354 | 0,360366 | -0,657916 | -0,314693 | -0,395506 | 0,159548 | -0,154296 | -0,040332 |
| |  | | --- | | X6 | | -0,385207 | 0,427123 | 0,170808 | -0,159803 | -0,133475 | 0,349741 | 0,159991 | 0,267655 | 0,614084 |
| |  | | --- | | X7 | | -0,449976 | -0,159297 | -0,150613 | -0,080807 | -0,154417 | 0,631150 | -0,171018 | -0,445596 | -0,304851 |
| |  | | --- | | X8 | | -0,296611 | 0,541065 | 0,169548 | -0,033681 | 0,284450 | -0,100477 | 0,279475 | 0,139828 | -0,633027 |
| |  | | --- | | X9 | | 0,296124 | -0,109345 | 0,345595 | -0,408014 | 0,692351 | 0,335770 | -0,090861 | -0,087288 | 0,080477 |

Следовательно, получаем матрицу коэффициентов линейного преобразования центрировано-нормированных исходных признаков (j = )

При снижении размерности признакового пространства до трех главных компонент следует рассматривать только три первых столбца матрицы U.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | Factor-variable correlations (factor loadings), based on correlations (Переделка) | | | | | | | | |
| |  | | --- | | Factor 1 | | |  | | --- | | Factor 2 | | |  | | --- | | Factor 3 | | |  | | --- | | Factor 4 | | |  | | --- | | Factor 5 | | |  | | --- | | Factor 6 | | |  | | --- | | Factor 7 | | |  | | --- | | Factor 8 | | |  | | --- | | Factor 9 | |
| Число дорожно-транспортных происшествий и пострадавших в них на 100 000 человек населения | **-0,303022** | **-0,754049** | **0,147043** | 0,298932 | 0,125874 | 0,117587 | 0,437056 | 0,088938 | 0,002516 |
| Смертность населения старше трудоспособного возраста, на 100 000 человек населения соответствующего возраста | **-0,744039** | **-0,440573** | **-0,095710** | -0,059986 | 0,178877 | -0,072118 | -0,325964 | 0,308915 | -0,025756 |
| Продажа сильно алкогольной продукции населению(тысяч декалитров)/**на тыс населения** | **-0,791890** | **0,037097** | **-0,115592** | 0,184008 | 0,399561 | -0,297659 | -0,027210 | -0,248219 | 0,117092 |
| Средняя Стоимость минимального (условного) набора потребительских товаров и услуг | **0,072830** | **0,002169** | **-0,862250** | -0,427294 | 0,139211 | -0,004756 | 0,216754 | 0,045819 | 0,012306 |
| Число спортивных сооружений/ **на тыс населения** | **-0,403272** | **-0,379990** | **0,390691** | -0,603697 | -0,274099 | -0,290117 | 0,102656 | -0,076542 | -0,013969 |
| Доходы консолидированных бюджетов субъектов Российской Федерации / **на тыс населения** | **-0,671885** | **0,583081** | **0,185182** | -0,146633 | -0,116257 | 0,256547 | 0,102941 | 0,132776 | 0,212687 |
| Предварительно расследовано преступлений, совершенных в состоянии алкогольного опьянения/ **на тыс населения** | **-0,784858** | **-0,217462** | **-0,163287** | -0,074148 | -0,134497 | 0,462969 | -0,110036 | -0,221047 | -0,105585 |
| Среднедушевые доходы населения (в месяц), руб. | **-0,517354** | **0,738626** | **0,183816** | -0,030905 | 0,247757 | -0,073704 | 0,179820 | 0,069365 | -0,219248 |
| Численность студентов, обучающихся по программам бакалавриата, специалитета, магистратуры на 10 000 человек населения, всего | **0,516506** | **-0,149271** | **0,374677** | -0,374389 | 0,603040 | 0,246298 | -0,058462 | -0,043301 | 0,027873 |

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 пятью главными компонентами наблюдается тесная связь Это подтверждает ошибочность выделения только трех первых главных компонент. Дадим интерпретацию главным признакам.

Поэтому согласно матрице нагрузок, выделим 5 ГК

ГК 1 :

имеет тесную отрицательную связь с Смертность населения старше трудоспособного возраста, на 100 000 человек населения соответствующего возраста; Продажа сильно алкогольной продукции населению(тысяч декалитров)/на тыс населения; Доходы консолидированных бюджетов субъектов Российской Федерации / на тыс населения; Предварительно расследовано преступлений, совершенных в состоянии алкогольного опьянения/ на тыс населения ее можно охарактеризовать как траты на алкоголь в регионе

ГК 2:

Имеет сильную отрицательную связь с Число дорожно-транспортных происшествий и пострадавших в них на 100 000 человек населения

И положительные связи с Доходы консолидированных бюджетов субъектов Российской Федерации / на тыс населения;и Среднедушевыми доходами населения этот компонент можно охарактеризовать как «Благосостояние региона»

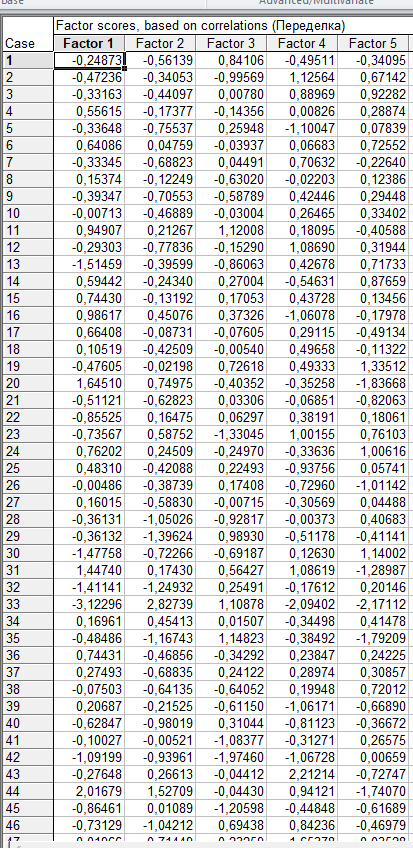
ГК 3:

Тесно отрицательно связана с Средняя Стоимость минимального (условного) набора потребительских товаров и услуг ее можно охарактеризовать как «дешевые услуги в регионе»

ГК 4: Имеет тесную отрицательную связь с количеством спортивных сооружений этот признак можно назвать «Неспортивность населения»

ГК 5: Имеет сильную положительную связь с числом студентов эту компоненту можно назвать образованность населения

Центрировано-нормированные исходные признаки связаны с центрировано-нормированными главными компонентами f1, f2, f3, f4, f5



3. Метод Уорда (МГК)

Методом Уорда при пороговом значении расстояния ρпор=12,46 все субъекты РФ разбиваются на 5 классов S1= {S11, S12, S13, S14, S15}, состав которых приведен в таблице 6.



|  |  |  |
| --- | --- | --- |
| **Номер кластера** | **Кол-во объектов** | **Состав класса** |
| Кластер 1 {S11} | 3 | Санкт-Петербург, Москва, Томская Область |
| Кластер 2 {S12} | 4 | Кабардино-Балкарская Республика, Республика Дагестан, Чеченская Республика, Республика Ингушетия |
| Кластер 3 {S13} | **6** | Ненецкий Автономный Округ, Камчатский Край, Магаданская Область, Ханты-Мансийский Автономный Округ, Чукотский Автономный Округ, Ямало-Ненецкий Автономный Округ |
| Кластер 4 {S14} | 28 | Республика Башкортостан, Смоленская Область, Республика Адыгея, Ростовская Область, Кемеровская область, Республика Мордовия, Липецкая Область, Республика Марий-Эл, Пензенская Область, Карачаево-Черкесская Республика, Амурская Область, Курганская Область, Оренбургская Область, Тверская Область, Новгородская Область, Чувашия, Ставропольский Край, Республика Северная Осетия-Алания, Республика Бурятия, Курская Область, Республика Хакассия, Омская Область, Белгородская Область, Тюменская Область, Тамбовская Область, Воронежская Область, Республика Калмыкия, Республика Тыва. |
| Кластер 5 {S15} | 44 | Вологодская Область, Приморский Край, Ульяновская Область, Самарская Область, Костромская Область, Красноярский Край, Псковская Область, Ивановская Область, Кировская Область, Республика Коми, Орловская Область, Астраханская Область, Волгоградская Область, Челябинская Область, Хабаровский Край, Архангельская Область, Мурманская Область, Новосибирская Область, Владимирская Область, Республика Карелия, Свердловская Область, Саратовская Область, Рязанская Область, Иркутская Область, Пермский Край, Еврейская Автономная Область, Ленинградская Область, Забайкальский Край, Сахалинская Область, Республика Татарстан, Удмуртская Республика, Краснодарский Край, Севастополь, Ярославская Область, Алтайский Край, Нижегородская Область, Тульская Область, Московская Область, Калининградская Область, Республика Саха, Калужская Область, Республика Крым, Республика Алтай, Брянская Область. |

4. Метод К-средних (МГК)

|  |  |  |
| --- | --- | --- |
| **Номер кластера** | **Кол-во объектов** | **Состав класса** |
| Кластер 1 {S11} | 42 | Вологодская Область, Приморский Край, Ульяновская Область, Самарская Область, Костромская Область, Красноярский Край, Псковская Область, Ивановская Область, Кировская Область, Республика Коми, Астраханская Область, Волгоградская Область, Челябинская Область, Хабаровский Край, Архангельская Область, Мурманская Область, Владимирская Область, Республика Карелия, Свердловская Область, Саратовская Область, Рязанская Область, Иркутская Область, Пермский Край, Ленинградская Область, Новгородская Область, Камчатский Край, Сахалинская Область, Курская Область, Удмуртская Республика, Краснодарский Край, Севастополь, Ярославская Область, Нижегородская Область, Тульская Область, Московская Область, Калининградская Область, Магаданская Область, Республика Саха, Республика Калмыкия, Калужская Область, Республика Крым, Республика Алтай. |
| Кластер 2 {S12} | 22 | Республика Башкортостан, Смоленская Область, Кемеровская Область, Республика Мордовия, Липецкая Область, Республика Марий-Эл, Пензенская Область, Амурская Область, Курганская Область, Оренбургская Область, Тверская Область, Еврейская Автономная Область, Забайкальский Край, Чувашская Республика, Республика Бурятия, Республика Хакассия, Алтайский Край, Белгородская Область, Тюменская Область, Тамбовская Область, Брянская Область, Республика Тыва |
| Кластер 3 {S13} | 9 | Орловская Область, Ростовская Область, Новосибирская Область, Республика Татарстан, Омская Область, Санкт-Петербург, Москва, Томская Область, Воронежская Область |
| Кластер 4 {S14} | 9 | Республика Адыгея, Кабардино-Балкарская Республика, Карачаево-Черкесская Республика, Республика Дагестан, Ставропольский Край, Республика Северная Осетия-Алания, Чеченская Республика, Ханты-Мансийский Автономный Округ, Республика Ингушетия |
| Кластер 5 {S15} | 3 | Ненецкий Автономный Округ, Чукотский Автономный Округ, Ямало-Ненецкий Автономный Округ |