**ФАКТОРНЫЙ АНАЛИЗ**

1. **Основные понятия**

**Факторный анализ** – раздел статистического многомерного анализа, объединяющий методы оценки размерности множества наблюдаемых переменных посредством исследования структуры ковариационных или корреляционных матриц.

Факторный анализ (ФА) представляет собой совокупность методов, которые на основе реально существующих связей анализируемых признаков, связей самих наблюдаемых объектов, позволяют выявлять скрытые (неявные, латентные) обобщающие характеристики организационной структуры и механизма развития изучаемых явлений, процессов. Основное предположение факторного анализа заключается в том, что корреляционные связи между большим количеством наблюдаемых переменных можно объяснить существованием меньшего числа гипотетических переменных или факторов, называемых общими скрытыми факторами. Далее их будем называть просто факторами.

Факторный анализ может быть:

* ***разведочным***— он осуществляется при исследовании скрытой факторной структуры без предположения о числе факторов и их нагрузках;
* ***конфирматорным***, предназначенным для проверки гипотез о числе факторов и их нагрузках.

**Фактор** – латентная переменная, конструируемая таким образом, чтобы можно было объяснить корреляцию между набором некоторых имеющихся переменных.

**Нагрузка** – корреляция между исходной переменной и фактором.

Общей моделью факторного анализа служит следующая линейная зависимость:

 (1)

где *Fj* – общие факторы, ;

*xi* – наблюдаемые переменные, ;

*Ui* – характерные факторы, ;

*εi* – случайные ошибки, .

Общую модель и математическую зависимость можно пояснить рис. 1.



Рис. 1 Схема модели факторного анализа.

Каждая переменная выражается как линейная комбинация латентных факторов:

 (2)

И наоборот, латентные факторы также можно выразить линейными комбинациями наблюдаемых переменных:

 (3)

Факторный анализ позволяет решить две важные проблемы исследователя:

* описать объект измерения ***всесторонне***, используя всю доступную информацию о нем: значения анализируемых признаков, связей между ними, а также связей между самими наблюдаемыми объектами;
* описать объект измерения ***компактно***, путем получения небольшого числа обобщающих факторов, объясняющих вариативность (дисперсию) элементарных признаков

Главными целями факторного анализа являются:

* сокращение числа переменных (редукция данных);
* определение структуры взаимосвязей между переменными, т.е. классификация переменных.

Поэтому факторный анализ используется или как метод сокращения данных, или как метод классификации.

При использовании методов факторного анализа решаются следующие задачи:

1. отыскание скрытых, но объективно существующих закономерностей исследуемого процесса, определяемых воздействием внутренних и внешних причин;
2. описание изучаемого процесса значительно меньшим числом факторов по сравнению с первоначально взятым количеством признаков;
3. выявление первоначальных признаков, наиболее тесно связанных с основными факторами;
4. прогнозирование процесса на основе уравнения регрессии, построенного по полученным факторам.
5. **Методика факторного анализа**

Практическое выполнение факторного анализа начинается с проверки его условий. В обязательные условия факторного анализа входят:

* все признаки должны быть количественными;
* число наблюдений должно быть не менее чем в два раза больше числа переменных;
* выборка должна быть однородна.

Для проведения факторного анализа предлагается методика, включающая в себя следующие этапы:

1. сбор исходных статистических данных и подготовка корреляционной (ковариационной) матрицы;
2. выделение общих скрытых факторов;
3. вращение факторной структуры;
4. содержательная интерпретация результатов факторного анализа.

Алгоритмы факторного анализа основываются на использовании редуцированной матрицы парных корреляций (ковариаций).

**Редуцированная матрица** – это матрица парных коэффициентов корреляции, на главной диагонали которой расположены не единицы (оценки) полной корреляции или оценки полной дисперсии, а их редуцированные, несколько уменьшенные величины – значения оценок **общностей**. Общность характеризует вклад данного признака в суммарную общность процесса

При этом постулируется, что в результате анализа будет объяснена не вся дисперсия изучаемых признаков (объектов), а ее некоторая часть, обычно большая. Оставшаяся необъясненная часть дисперсии — это характерность, возникающая из-за специфичности наблюдаемых объектов, или ошибок, допускаемых при регистрации явлений, процессов, т.е. ненадежности вводных данных.

1. **Выделение факторов**

Выделение общих скрытых факторов – это, прежде всего, выбор метода факторного анализа. Наиболее часто на практике используют методы: главных компонент (компонентный анализ) и главных факторов (факторный анализ).

Особняком в методах факторного анализа стоит метод главных компонент (МГК). Его главное отличие заключается в том, что обработке подлежит не редуцированная, а обычная матрица парных корреляций (ковариаций) на главной диагонали которой расположены единицы (в матрице ковариаций – оценки полной дисперсии). Иными словами, предполагается объяснение всей дисперсии анализируемых признаков (необходимых объектов), а явление «характерности» во внимание не принимается. Таким образом, коррелированные компоненты заменяются некоррелированными факторами.

Другой важной характеристикой метода является возможность ограничиться наиболее информативными главными компонентами и исключить остальные из анализа, что упрощает интерпретацию результатов.

Важным достоинство МГК является то, что он — единственный математически обоснованный метод факторного анализа. По утверждению ряда исследователей МГК не является методом факторного анализа, поскольку не расщепляет дисперсию индикаторов на общую и уникальную. Несмотря на это, МГК имеет схожий алгоритм и решает схожие аналитические задачи.

В большинстве случаев эти два метода приводят к весьма близким результатам. Компонентный анализ более предпочтителен как метод сокращения данных, а факторный анализ лучше применять с целью определения структуры данных.

При классификации методов ФА можно выделить следующие группы:

1. ***Метод главных компонент***;
2. *Упрощенные методы ФА,* обычно это методы, которые появились раньше, в первой половине двадцатого столетия, во время появления и формирования базисных теоретических разработок ФА. Эти методы отличаются, с одной стороны, сравнительно простыми вычислительными процедурами, а с другой стороны, ограниченными возможностями в выделении латентных факторов и аппроксимации факторных решений. В данную группу входят методы:

* ***однофакторная модель*** (Ч. Спирмен)*,* позволяет выделять только один латентный фактор;
* ***бифакторная модель*** (Г. Хользингер)*,* ориентирована на выделение двух латентных факторов;

1. *Современные аппроксимирующие методы ФА* – методы, имеющие, по сравнению с предыдущей группой, более гибкую модель выделения латентных факторов (искусственно не ограничивающую их число), а также позволяющую оптимизировать полученные решения. В этой группе наиболее представительными являются:

* ***метод главных факторов*** (Г. Томсон) используется на практике особенно часто, наиболее близок методу главных компонентов;
* ***групповой метод*** (Л. Гуттман и П. Хорст) основывается на исследовании не простого набора данных, а на предварительно отобранных каких-либо группах анализируемых признаков (наблюдаемых объектов).

1. *Методы с повышенными аппроксимирующими свойствами* – современные методы, позволяющие получать и последовательно улучшать аналитические результаты. Эти методы отличаются сложностью алгоритмов и высокой трудоемкостью вычислительных процедур, практически нереализуемы без технических средств. К этой группе относятся методы:

* ***максимального правдоподобия*** (Д. Лоули и Д. Максвелл);
* ***минимальных остатков*** (Г. Харман);
* ***двухфакторного анализа*** (Г. Кайзер и И. Кэффри);
* ***канонического факторного анализа*** (К. Рао).

Проведем факторный анализ методом главных компонент населения муниципальных образований РТ (по данным ежегодного статистического сборника) в ППП IBM SPSS Statistics v20.

|  |  |  |
| --- | --- | --- |
| Таблица 1. Общности | | |
|  | Начальные | Извлеченные |
| моложе трудоспособного возраста оба пола (%) | 1,000 | ,944 |
| моложе трудоспособного возраста мужчины (%) | 1,000 | ,942 |
| моложе трудоспособного возраста женщины (%) | 1,000 | ,896 |
| трудоспособного возраста оба пола (%) | 1,000 | ,992 |
| трудоспособного возраста мужчины (%) | 1,000 | ,957 |
| трудоспособного возраста женщины (%) | 1,000 | ,970 |
| старше трудоспособного возраста оба пола (%) | 1,000 | ,988 |
| старше трудоспособного мужчины (%) | 1,000 | ,971 |
| старше трудоспособного женщины (%) | 1,000 | ,974 |
| число родившихся на 1000 человек | 1,000 | ,652 |
| число умерших на 1000 человек | 1,000 | ,892 |
| число умерших детей до 1 года на 1000 родившихся | 1,000 | ,660 |
| естественный прирост (убыль) на 1000 человек | 1,000 | ,891 |
| число зарегистрированных браков на 1000 человек | 1,000 | ,827 |
| число разводов на 1000 человек | 1,000 | ,666 |
| среднесписочная численность работающих | 1,000 | ,545 |
| численность зарегистрированных безработных | 1,000 | ,785 |
| численность пенсионеров | 1,000 | ,792 |
| численность пенсионеров на 1000 человек | 1,000 | ,617 |
|  | | |

В табл. 1 показано, какую часть дисперсии каждой из включенных в анализ переменных объясняет факторная модель. Если значение в столбце «Извлеченные» невелико, переменную следует исключить из анализа. Значения в столбце «Начальные» равны единице, что соответствует диагонали используемой корреляционной матрицы.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Таблица 2. Факторная нагрузка | | | | | | | | | |
| Компонента | Начальные собственные значения | | | Суммы квадратов нагрузок извлечения | | | Суммы квадратов нагрузок вращения | | |
| Итого | % Дисперсии | Кумулятивный % | Итого | % Дисперсии | Кумулятивный % | Итого | % Дисперсии | Кумулятивный % |
| 1 | 7,086 | 37,297 | 37,297 | 7,086 | 37,297 | 37,297 | 6,532 | 34,378 | 34,378 |
| 2 | 4,576 | 24,086 | 61,383 | 4,576 | 24,086 | 61,383 | 4,521 | 23,795 | 58,173 |
| 3 | 1,749 | 9,205 | 70,589 | 1,749 | 9,205 | 70,589 | 1,997 | 10,509 | 68,682 |
| 4 | 1,504 | 7,913 | 78,502 | 1,504 | 7,913 | 78,502 | 1,523 | 8,015 | 76,697 |
| 5 | 1,044 | 5,497 | 83,999 | 1,044 | 5,497 | 83,999 | 1,387 | 7,302 | 83,999 |
| 6 | ,969 | 5,098 | 89,096 |  |  |  |  |  |  |
| 7 | ,669 | 3,519 | 92,615 |  |  |  |  |  |  |
|  | | | | | | | | | |

Во втором столбце табл. 2 указываются дисперсии выделенных общих факторов. В третьем столбце для каждого общего фактора приводится процент от общей дисперсии. Как можно видеть, первый фактор объясняет 37,3% общей дисперсии, фактор 2 – 24,1%, и т.д. по убыванию. Четвертый столбец содержит накопленный процент от общей дисперсии.

Дисперсии, выделяемых общих факторов, называют **собственными значениями**. Это название происходит от использованного способа вычисления.

Как только получена информация о том, сколько дисперсии выделил каждый фактор, необходимо определить количество факторов. По своей природе это решение произвольно. Однако имеются некоторые общеупотребительные рекомендации, и на практике следование им дает наилучшие результаты.

**Критерий Кайзера.** Сначала вы можете отобрать только факторы, с собственными значениями, большими 1. По существу, это означает, что если фактор не выделяет дисперсию, эквивалентную, по крайней мере, дисперсии одной переменной, то он опускается. Этот критерий предложен Кайзером (Kaiser,1960), и является, вероятно, наиболее широко используемым. В приведенном выше примере на основе этого критерия следует сохранить только 5 факторов (пять главных компонент).

**Критерий каменистой осыпи** является графическим методом, впервые предложенным Кэттелем. Собственные значения, представляются в виде простого графика. Кэттель предложил найти такое место на графике, где убывание собственных значений слева направо максимально замедляется. Предполагается, что справа от этой точки находится только "факториальная осыпь".

В результате выполнения процедуры для примера, получаем график собственных значений общих факторов (рис. 3). При помощи критерия Кайзера отбираем 3 общих фактора.



Рис. 3. График собственных значений общих факторов.

Как видно, выделение трех факторов методом каменистой осыпи недостаточно, т.к. ими объясняется лишь 70,6% дисперсии, в то время как 5 факторов, выделенных критерием Кайзера, объясняется 84%.

Таким образом, выделением пяти общих факторов удалось объяснить 84% общей дисперсии, что вполне достаточно. Для дальнейшего исследования оставим пять факторов, перечисленных в таблице 2.

Результатом выполнения второго этапа факторного анализа (выделения общих ортогональных факторов) является таблица факторных нагрузок (таблица 3). Факторные нагрузки представляют собой коэффициенты корреляции между выделяемыми общими факторами *Fj*, и производственно-экономическими факторами *xi*, .



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Таблица 3. Факторные нагрузки | | | | | |
|  | Компонента | | | | |
| 1 | 2 | 3 | 4 | 5 |
| моложе трудоспособного возраста оба пола (%) | ,260 | ,915 | -,121 | ,144 | ,057 |
| моложе трудоспособного возраста мужчины (%) | ,240 | ,903 | -,185 | ,163 | ,095 |
| моложе трудоспособного возраста женщины (%) | ,259 | ,899 | -,059 | ,130 | ,029 |
| трудоспособного возраста оба пола (%) | ,855 | -,503 | -,043 | ,077 | -,010 |
| трудоспособного возраста мужчины (%) | ,736 | -,641 | ,016 | ,047 | ,049 |
| трудоспособного возраста женщины (%) | ,888 | -,404 | -,089 | ,095 | -,037 |
| старше трудоспособного возраста оба пола (%) | -,977 | ,027 | ,105 | -,150 | -,019 |
| старше трудоспособного мужчины (%) | -,955 | ,099 | ,104 | -,159 | -,115 |
| старше трудоспособного женщины (%) | -,969 | -,001 | ,112 | -,148 | ,023 |
| число родившихся на 1000 человек | ,225 | ,377 | ,670 | ,105 | -,020 |
| число умерших на 1000 человек | -,731 | -,476 | ,203 | ,179 | ,241 |
| число умерших детей до 1 года на 1000 родившихся | -,234 | -,227 | ,067 | ,739 | -,046 |
| естественный прирост (убыль) на 1000 человек | ,717 | ,557 | ,086 | -,113 | -,215 |
| число зарегистрированных браков на 1000 человек | ,163 | ,425 | ,204 | -,522 | ,553 |
| числов разводов на 1000 человек | ,321 | -,583 | -,012 | -,238 | ,406 |
| среднесписочная численность работающих | ,676 | -,030 | -,046 | -,187 | ,223 |
| численность зарегистрированных безработных | -,200 | ,156 | ,137 | ,583 | ,602 |
| численность пенсионеров | ,456 | -,043 | ,742 | ,140 | -,113 |
| численность пенсионеров на 1000 человек | ,201 | -,004 | ,734 | -,164 | -,106 |

1. **Вращение**

Следующим этапом является “поворот” факторов.

Задача вращения общих факторов решается с целью улучшения их интерпретируемости. Производится попытка достижения простой структуры, в которой каждая переменная характеризуется преобладающим влиянием какого–то одного фактора. Факторные нагрузки могут быть изображены в виде диаграммы рассеяния, на которой каждая переменная представлена точкой. Можно повернуть оси в любом направлении без изменения относительного положения точек. При этом действительные координаты точек, то есть факторные нагрузки, изменяются. Наглядно это изображено на рис. 4.



Рис. 4 вращение факторов

Существуют различные методы вращения факторов. Целью этих методов является получение понятной (интерпретируемой) матрицы нагрузок, то есть факторов, которые ясно отмечены высокими нагрузками для некоторых переменных и низкими - для других.

Примеры методов вращения:

* ***Варимакс***. Ортогональный метод вращения, минимизирующий число переменных с высокими нагрузками на каждый фактор.
* ***Метод косоугольного (неортогонального) вращения***. Самое косоугольное решение соответствует дельте, равной 0 (по умолчанию). По мере того, как дельта отклоняется в отрицательную сторону, факторы становятся более ортогональными.
* ***Квартимакс***. Метод вращения, который минимизирует число факторов, необходимых для объяснения каждой переменной.
* ***Эквимакс***. Метод вращения, объединяющий методы варимакс, упрощающий факторы, и квартимакс, упрощающий переменные. Минимизируется число переменных с большими факторными нагрузками и число факторов, требуемых для объяснения переменной.
* ***Промакс-вращение***. Косоугольное вращение в предположении, что факторы могут коррелировать между собой. Оно производится быстрее, чем вращение типа косоугольного вращения, поэтому оно полезно для больших наборов данных.
* и т.д.

В исходном примере целесообразно применение метода вращения *Варимакс*, максимизирующего разброс квадратов факторных нагрузок по каждому фактору в отдельности и приводящий к увеличению больших нагрузок и уменьшению маленьких нагрузок.

Результаты вращения представлены в табл. 4.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Таблица 4. Матрица повернутых компонент | | | | | |
|  | Компонента | | | | |
| 1 | 2 | 3 | 4 | 5 |
| моложе трудоспособного возраста оба пола (%) | -,001 | ,958 | ,022 | ,134 | ,089 |
| моложе трудоспособного возраста мужчины (%) | ,002 | ,952 | -,050 | ,135 | ,122 |
| моложе трудоспособного возраста женщины (%) | -,011 | ,932 | ,083 | ,128 | ,067 |
| трудоспособного возраста оба пола (%) | ,961 | -,217 | ,095 | -,057 | -,100 |
| трудоспособного возраста мужчины (%) | ,886 | -,397 | ,103 | -,035 | -,050 |
| трудоспособного возраста женщины (%) | ,968 | -,103 | ,072 | -,067 | -,118 |
| старше трудоспособного возраста оба пола (%) | -,947 | -,277 | -,104 | -,013 | ,053 |
| старше трудоспособного мужчины (%) | -,960 | -,198 | -,078 | -,052 | -,026 |
| старше трудоспособного женщины (%) | -,928 | -,306 | -,106 | ,009 | ,085 |
| число родившихся на 1000 человек | -,004 | ,314 | ,732 | ,066 | ,117 |
| число умерших на 1000 человек | -,507 | -,656 | -,041 | -,145 | ,427 |
| число умерших детей до 1 года на 1000 родившихся | -,052 | -,136 | ,018 | -,636 | ,484 |
| естественный прирост (убыль) на 1000 человек | ,435 | ,687 | ,320 | ,151 | -,322 |
| число зарегистрированных браков на 1000 человек | -,036 | ,248 | ,170 | ,854 | ,081 |
| числов разводов на 1000 человек | ,485 | -,537 | -,071 | ,359 | ,091 |
| среднесписочная численность работающих | ,634 | ,101 | ,056 | ,355 | -,066 |
| численность зарегистрированных безработных | -,111 | ,123 | ,021 | -,022 | ,870 |
| численность пенсионеров | ,325 | -,010 | ,825 | -,067 | ,041 |
| численность пенсионеров на 1000 человек | ,035 | -,098 | ,759 | ,134 | -,110 |

После проведения вращения методом *Варимакс*, факторная структура существенно изменилась. К положительным результатам отнесем то, что факторов стали ясно отмечены высокими нагрузками для некоторых переменных и низкими - для других. Красным цветом в табл. 4 выделены переменные, вошедшие в соответствующий фактор.

Следующим этапом является проведение интерпретации получившихся результатов.

1. **Интерпретация результатов**

Главной проблемой факторного анализа является выделение и интерпретация главных факторов. При отборе компонент исследователь обычно сталкивается с существенными трудностями, так как не существует однозначного критерия выделения факторов, и потому здесь неизбежен субъективизм интерпретаций результатов.

Приведем интерпретацию получившихся результатов и дадим название выделенным факторам.

*F1*: В данный общий фактор в качестве существенных входят следующие исходные переменные*: x4*– процент населения трудоспособного возраста; *x5*– процент мужского населения трудоспособного возраста; *x6*– процент женского населения трудоспособного возраста; *x7*– процент населения старше трудоспособного возраста; *x8*– процент мужского населения старше трудоспособного возраста; *x9*– процент женского населения старше трудоспособного возраста; *x16*– среднесписочная численность работающих. Этот общий фактор включает в себя переменные, характеризующие трудовые ресурсы населения, поэтому назовем его трудовыми ресурсами.

*F2*: В данный общий фактор в качестве существенных входят следующие исходные переменные*: x1*– процент населения моложе трудоспособного возраста; *x2*– процент мужского населения моложе трудоспособного возраста; *x3*– процент женского населения моложе трудоспособного возраста; *x11*– число умерших на 1000 человек; *x13*– естественный прирост (убыль) на 1000 человек; *x15*– числов разводов на 1000 человек. Этот общий фактор связан положительно с переменными характеризующими население моложе трудового возраста, естественным приростом населения и отрицательно с переменными, характеризующими число умерших и разводов. Назовем этот фактор потенциалом воспроизводства трудовых ресурсов.

*F3*: В данный общий фактор в качестве существенных входят следующие исходные переменные*: x10*– число родившихся на 1000 человек; *x18*– численность пенсионеров; *x19*– численность пенсионеров на 1000 человек. Этот общий фактор связан положительно с переменными характеризующими численность родившихся и численность пенсионеров. Назовем этот фактор новорожденные и пенсионеры.

*F4*: В данный общий фактор в качестве существенных входят следующие исходные переменные*: x12*– число умерших детей до 1 года на 1000 родившихся; *x14*– число зарегистрированных браков. Этот общий фактор связан положительно с переменными характеризующими число умерших детей в возрасте до 1 года и число зарегистрированных браков. Назовем этот фактор уровень зависимости младенческой смертности от числа зарегистрированных браков.

*F5*: В данный общий фактор в качестве существенных входят следующие исходные переменные*: x17*– численность зарегистрированных безработных. Назовем этот фактор численность зарегистрированных безработных.