**Лабораторная работа №6.**

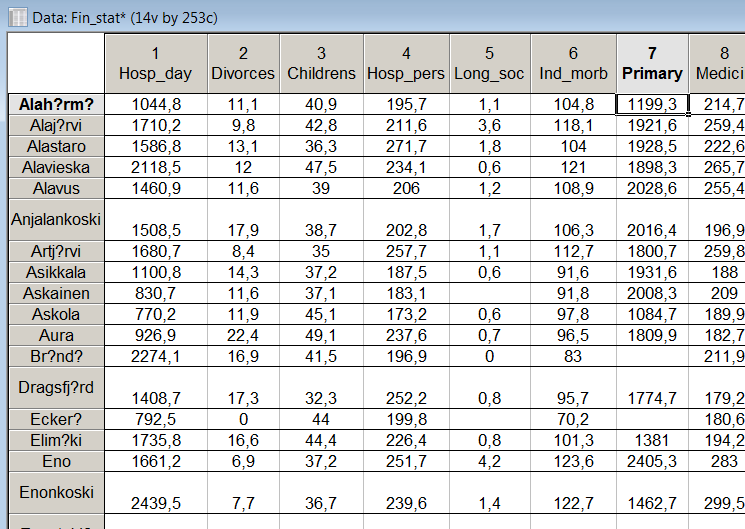
Тема «Компонентный анализ» (**время выполнения 2 часа**).

**Срок сдачи до 06.05.13**

Провести компонентный анализ исходных признаков X1, X2, X3, X4, X5 (из лабораторной №4). Определить необходимое число компонент. Вычислить факторные нагрузки. Определить вклады факторов в суммарную дисперсию. Интерпретировать найденные главные компоненты с использованием факторных нагрузок и их квадратов. Определить, существует ли связь между первым фактором и признаком Y (результирующий признак из лабораторной работы №4).

1. **Вычисление компонент**

В качестве примера рассмотрим значения некоторых медико-социальных показателей регионов Финляндии (более 50 регионов):

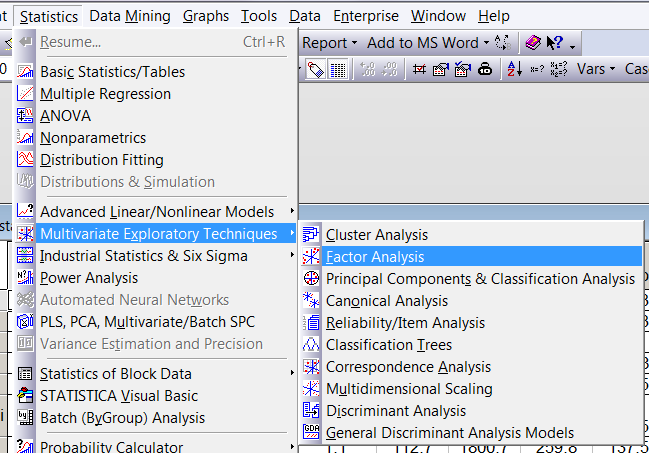


Для поиска компонент будем использовать 1-6 признаки:

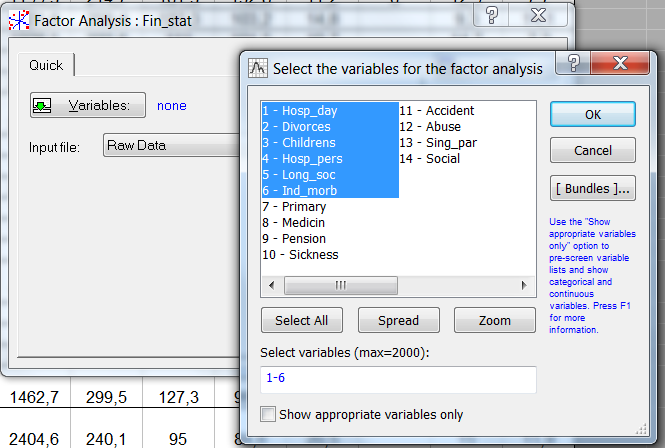
|  |  |
| --- | --- |
| Hosp\_day | Кол-во койко-дней в стационаре на 1000 чел населения |
| Divorces | Количество разводов на 1000 состоящих в браке |
| Childrens | % семей с детьми |
| Hosp\_pers | Кол-во пролечившихся в стационаре на 1000 чел |
| Long\_soc | Количество потребителей долговременной социальной помощи в % ко всем |
| Ind\_morb | Индекс заболеваемости |

Перед применением процедуры компонентного анализа следует построить корреляционную матрицу для исследуемых признаков и проанализировать взаимосвязи (не строить графики, только словесное пояснение).

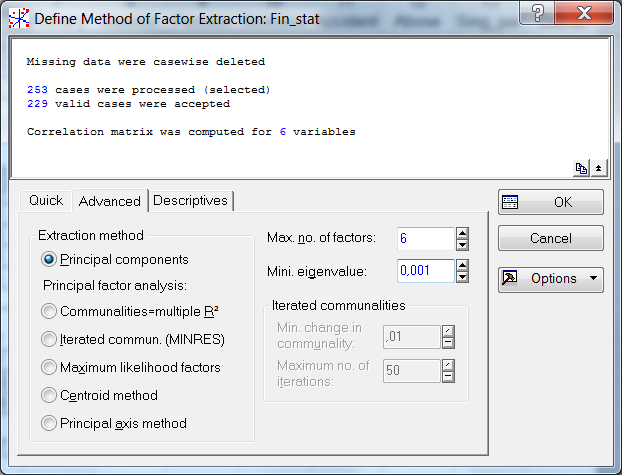
Для вызова модуля компонентного анализа и используется пункт меню «Statistics/Multivariate exploratory techniques/ Factor Analysis» (Многомерные разведочные методы/ Факторный анализ):

****

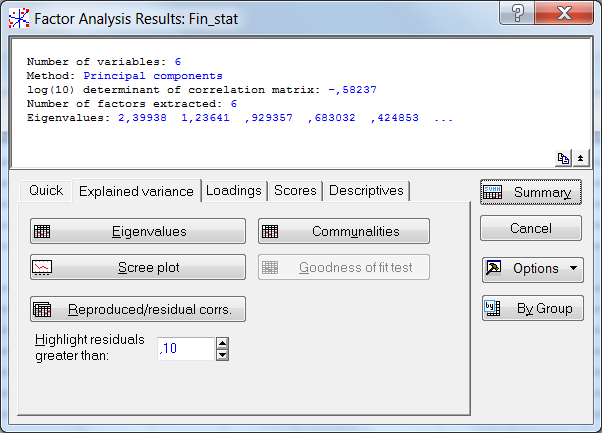
В открывшейся форме следует выбрать признаки, по которым будем проводить поиск главных компонент:



В следующей форме следует указать количество компонент, которые хотим построить по исходным данным («Max. no of factors») и минимальное значение собственного числа – собственные числа меньшие этого значения (и соответствующие им компоненты) будут исключаться из результатов («Mini. Eigenvalue»). Способ получения компонент – «Principal components» (Главные компоненты). На первом этапе исследования желательно указывать количество компонент равное количеству исследуемых признаков и достаточно малое значение минимального собственного числа:



Следующая форма используется для исследования полученных компонент:



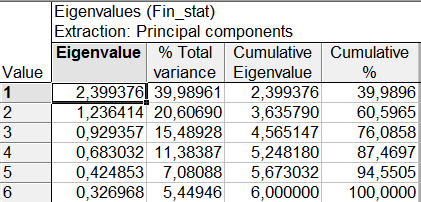
Для решения вопроса о количестве собственных чисел или количестве компонент «взявших на себя» наибольшую часть суммарной дисперсии исходных признаков можно рассмотреть график изменения значений собственных чисел (закладка «Explained variance», кнопка «Scree plot»):

 .

Существует несколько подходов к определению достаточного количества компонент с использованием этого графика:

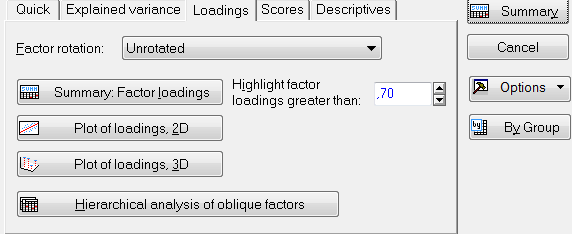
* Критерий Кайзера: оставляются компоненты, для которых . В данном примере – это первая вторая компоненты.
* Критерий «каменистой осыпи»: на графике находят место, где убывание значений слева направо максимально замедляется. Значения слева от этого участка – оставляем, справа – исключаем. В данном примере участок графика после второго собственного числа можно считать участком «замедления».
* По количеству факторов, которым можно придать смысл при интерпретации.

После вывода графика по нажатию на кнопку «Eigenvalues» можно получить следующую таблицу:

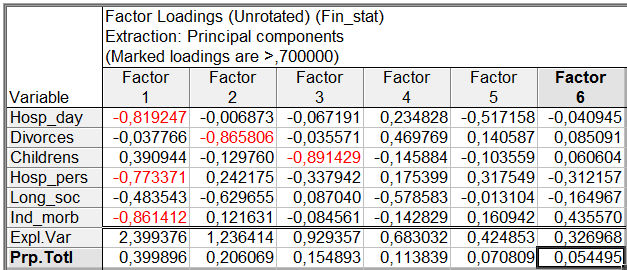


В ней содержатся: собственные значения, процент суммарной дисперсии, приходящейся на компоненту, соответствующую данному собственному числу (суммарная дисперсия равна количеству исходных признаков – следу корреляционной матрицы, см. лекции). Также в таблице представлены собственные числа и процент дисперсии с накоплением. По сумме накопленных процентов видно, что первые три компоненты объясняют более 76% суммарной дисперсии исходных признаков.

Для интерпретации факторов необходимо вычислить факторные нагрузки и их квадраты. На закладке «Loadings» по нажатию кнопки «Summary: Factor loadings» получаем значения факторных нагрузок.



Значения выводятся в виде таблицы:



Значения квадратов факторных нагрузок в системе STATISTICA можно получить только для первого фактора по нажатию кнопки «Communality» (закладка «Explained variance»). Можно вычислить квадраты в других программных системах и представить их в виде таблицы, можно для интерпретации факторов воспользоваться только значениями факторных нагрузок.

В данном примере наибольшие факторные нагрузки в первой компоненте падают на признаки «Hosp\_day», «Hosp\_pers» и «Ind\_morb», связанные с заболеваемостью и лечением в стационаре. Все признаки входят в компоненту с отрицательными знаками, поэтому высокие значения этих признаков будут давать низкие (отрицательные) значения компоненты. На основе этого можно сделать вывод, что данная компонента характеризует уровень заболеваемости и количество оказываемой медицинской помощи в регионе.

Вторая компонента имеет наибольший вклад признаков «Divorces» (разводы) и «Long\_soc» (долговременная соц.помощь). На основании этих признаков и равенства их знаков можно сделать вывод о том, что данная компонента характеризует количество социальных выплат неполным семьям.

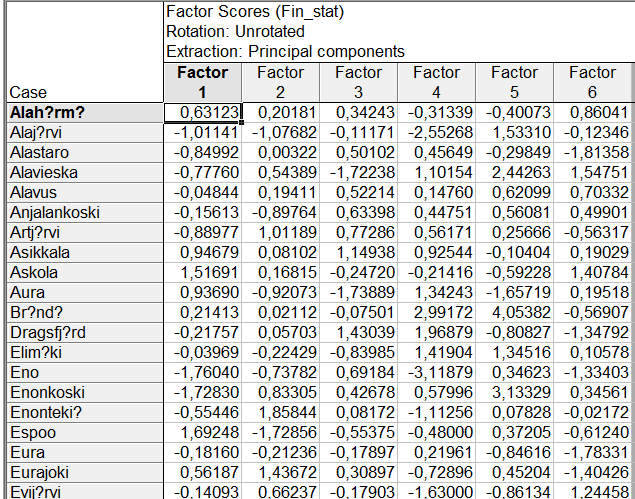
Третья компонента дублирует исходный признак «Childrens» и как комбинация нескольких исходных признаков интерпретироваться не может.

В интерпретации найденных компонент может помочь также диаграмма рассеивания исходных признаков в пространстве компонент (закладка «Loadings», кнопка «Plot of loadings 2D»):

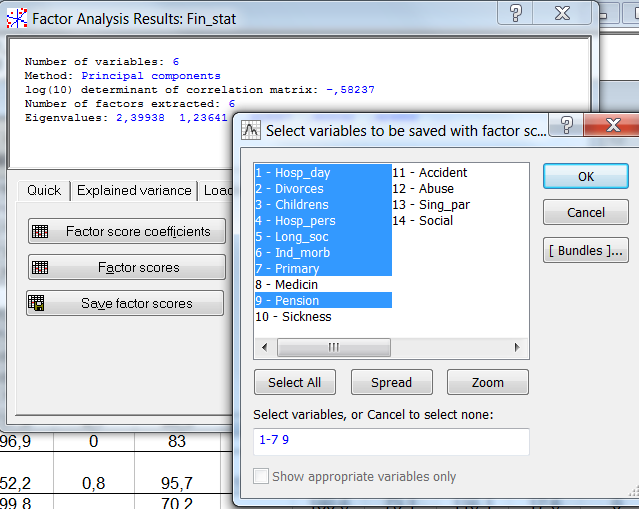


1. **Вычисление значений компонент для наблюдений.**

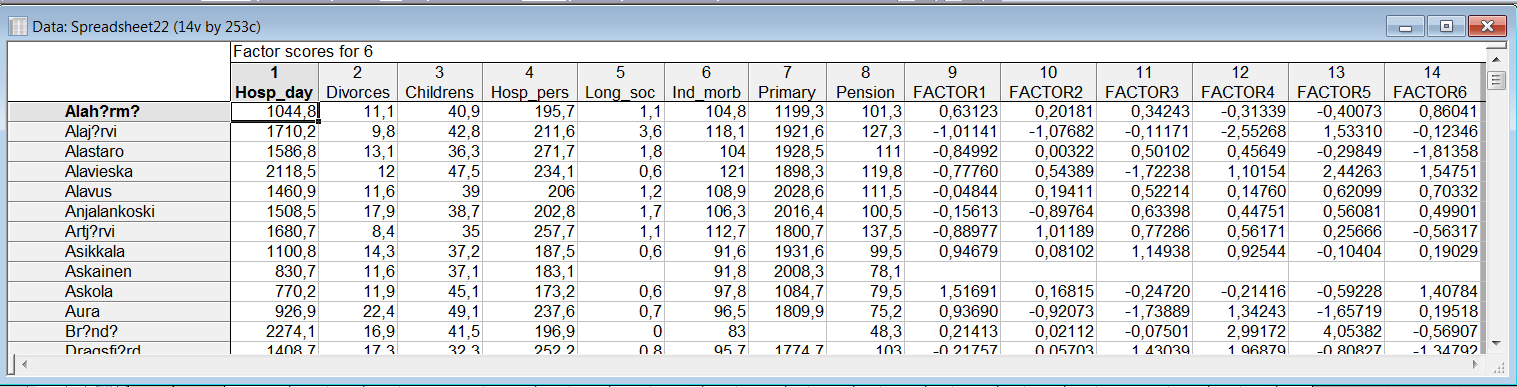
Вычислить значения компонент для каждого наблюдения можно с использованием кнопки «Factor Scores» (закладка «Scores»). Полученная таблица содержит значения компонент для каждого исходного наблюдения:



Полученные значения компонент можно сохранить вместе с исходными признаками (кнопка «Save Factor Scores», закладка «Scores»), указав исходные признаки, которые хотим сохранить:



В результате получим таблицу с исходными признаками и значениями компонент (новых признаков) для каждого наблюдения:



Данную таблицу можно использовать для дальнейших исследований, в том числе для расчета корреляций между признаками, по которым не строились компоненты, и компонентами (FACTOR1, FACTOR2 и т.д.)