Домашнее задание по МГК

Агалян Роберт

1

Загрузка библиотек и данных, очистка пропущенных значений

```
library(tidyverse)
pro <- read.csv("protests.csv")
d <- pro %>% dplyr::select(country, year, duration, part2,
reaction, arrested, wounded)
d <- na.omit(d)
```

Отбор и аггрегация:

```
final <- d %>% group_by(country, year) %>%
              summarise(duration = sum(duration),
                        part2 = sum(part2),
                        reaction = sum(reaction),
                        arrested = sum(arrested),
                        wounded = sum(wounded))
```

```
## `summarise()` has grouped output by 'country'. You can override using the
## `.groups` argument.
```

```
final <- as.data.frame(final)
m <- final[3:7]
```

2

Реализация метода гланвых компонент

```
pca <- prcomp(m, center = TRUE, scale = TRUE)
pca
```

```
## Standard deviations (1, .., p=5):
## [1] 1.9294196 0.7419670 0.6484467 0.4713142 0.2901804
##
## Rotation (n x k) = (5 x 5):
##                 PC1        PC2        PC3         PC4         PC5
## duration -0.4603603  0.4646695 -0.3527759  0.12505015  0.65731445
## part2    -0.4183310  0.1693309  0.8884696 -0.04182172  0.07210334
## reaction -0.4748435  0.4033850 -0.2408012 -0.01540626 -0.74403065
## arrested -0.4345467 -0.5955446 -0.0561144  0.67204217 -0.04130654
## wounded  -0.4458255 -0.4878692 -0.1582305 -0.72851587  0.08631902
```

Выражение для второй главной компоненты:

$$\mathrm{PC}_2 = 0.465 \times \mathrm{duration} + 0.169 \times \mathrm{part2} + 0.403 \times \mathrm{reaction} - 0.596 \times \mathrm{arrested} - 0.488 \times \mathrm{wounded}$$

3

Интерпретация

Интерпретация первой главной компоненты: В первую ГК с приблизительно одинаковыми отрицательными весами входят все переменные. Можем предложить назвать этот индикатор “влиятельность протеста” или “градус протеста”. То есть, предполагается, что этот индикатор показывает и влияние протеста на других людей, то есть его популярность, так и на госструктуры и различные реакции с их стороны.

Интерпретация второй главной компоненты: Вторая ГК включает в себя с положительными весами продолжительность протеста в днях, число его участников и реакцию на протест, это те характеристики, которые показывают распространённость протеста. С отрицательными весами входят число арестованных и раненых.

4

Выбор количества главных компонент

Используя правило Кайзера, мы можем выбрать столько ГК, сколько собственных значений больше 1, таких значенйи только одно, значит по этому правилу мы должны выбрать одну ГК.

Второй способ: выберем столько главных компонент, чтобы они объясняли не менее 75-80% дисперсии исходных данных:

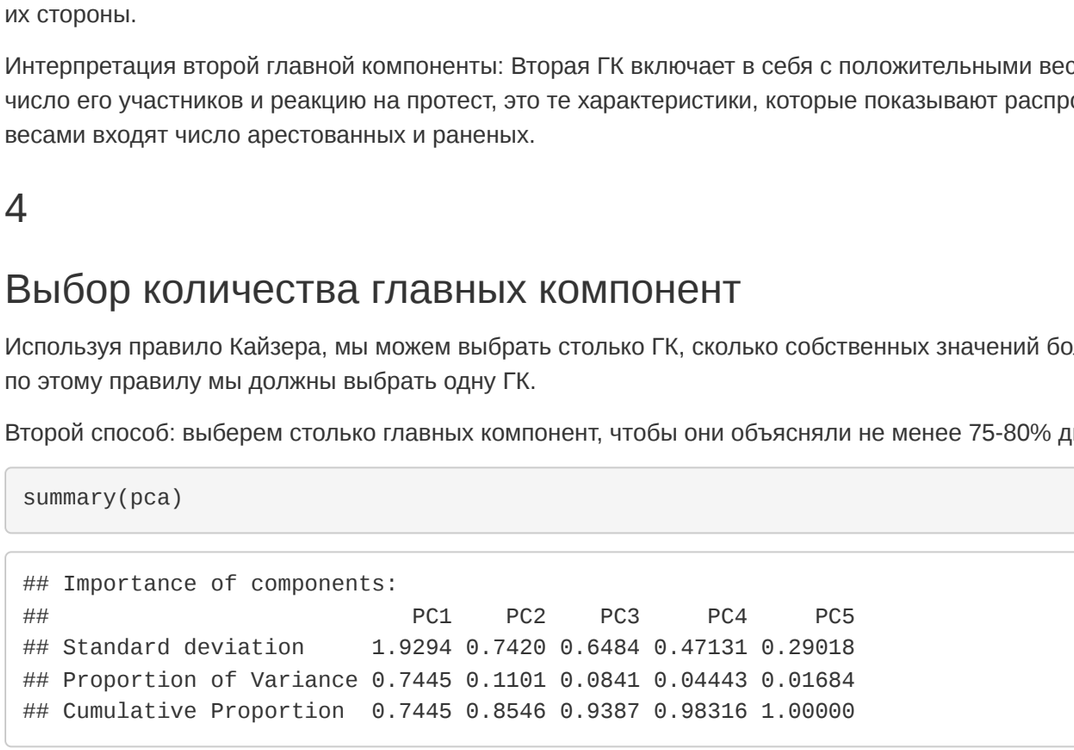
```
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5
## Standard deviation     1.9294 0.7420 0.6484 0.47131 0.29018
## Proportion of Variance 0.7445 0.1101 0.0841 0.04443 0.01684
## Cumulative Proportion  0.7445 0.8546 0.9387 0.98316 1.00000
```

Согласно этому критерию, стоит выбрать две ГК, так как вместе они объясняют примерно 85% дисперсии исходных данных.

Третий способ: используем правило Кэттела: выберем столько гланвных компонент, сколько наблюдается до излома на графике «каменистой осыпи»:

```
plot(pca, type = "l", main = "Scree plot")
```



Излом на графике наблюдается при количестве ГК, равном двум. Исходя из этого графика мы должны выбрать две ГК.

Исходя из трёх проведённых методов определения оптимального числа ГК, вероятно, стоит выбрать две ГК. Я ожидал такое количество, однако предполагал, что первая компонента будет интерпретироваться как “мощность протеста” в виде числа его участников и его продолжительность, а вторая - “реакция государства”, состоящая из реакции, числа арестов и раненых.