HDDA Tutorial: FactorModel : Solutions

Department of Econometrics and Business Statistics, Monash University

Tutorial 8

We will investigate the Boston housing data also covered in the lecture.

Factor analysis

Carry out factor analysis using a **four** factor model. At first use no rotation.

```
#First load required packages
library(tidyverse)
Boston<-readRDS('Boston.rds')
Boston%>%
  column_to_rownames('Town')%>%
  factanal(factors = 4,rotation = 'none',scores = 'none')->fa
```

1. What is the loading of the third factor on the variable PTRATIO

```
loadings(fa)["PTRATIO",3]
```

```
## [1] -0.6802375
```

2. What are the unique variance of the variable MEDV and CRIM?

```
fa$uniquenesses["MEDV"]
```

```
##     MEDV
## 0.126886
```

```
fa$uniquenesses["CRIM"]
```

```
##      CRIM
## 0.7086018
```

3. Carry out the analysis after standardising the data first. Does your answer to question 2 change?

```
Boston%>%
  column_to_rownames('Town')%>%
  scale%>%
  factanal(factors = 4,rotation = 'none',scores = 'none')->fa_sc
fa_sc$uniquenesses["MEDV"]
```

```
##     MEDV
## 0.126886
```

```
fa_sc$uniquenesses["CRIM"]
```

```
##      CRIM
## 0.7086018
```

```
# The answer is the same
```

4. Why are the answers to 3 and 4 the same/ different?

The asnswers are the same since the R function automatically standardises the data. This also implies that the total variance of each variable is 1.

5. In light of all these answers, interpret the uniqueness of MEDV and CRIM and compare these.

The uniqueness of MEDV is 0.1269 *which implies that* 12.69% *of the variation in median house value cannot be explained by factors common to all variables in the analysis.*

The uniqueness of CRIM is 0.7086 *which implies that* 70.86% *of the variation in crime cannot be explained by factors common to all variables in the analysis.*

Compared to median house value, a much larger component of crime is idiosyncratic and is not explained by factors common to all variables in the analysis

Interpreting the factors

1. For each factor identify the loadings that are close to zero and those that are large.

All variables apart from CHAS load onto the first factor with the largest loadings for RAD and TX. All variables apart from CHAS, CRIM and TX load onto the second factor with the largest loading for NOX. Factors 3 and 4 have similar patterns of loadings.

2. Carry out the same analysis after doing a varimax rotation.

```
Boston%>%
  column_to_rownames('Town')%>%
  factanal(factors = 4,rotation = 'varimax',scores = 'none')->fa_v
```

The Varimax rotation has not worked too well since now the first three factors have very few zero loadings. This is also difficult to interpret.

3. Carry out the same analysis after doing a varimax rotation.

```
Boston%>%
  column_to_rownames('Town')%>%
  factanal(factors = 4,rotation = 'promax',scores = 'none')->fa_p
```

The first factor has high loadings for AGE and DIS, these have almost zero loadings for the other factors. The loading for ZN is also quite high. The first factor is a ***geographic factor*** *with higher scores associated with older houses close to the city.*

The second factor has high loadings for RAD and TX, these have almost zero loadings for the other factors. The second factor is also the only factor that CRIM loads onto. The second factor is a ***infrastructure factor*** *with higher scores associated with higher property taxes and better access to highways.*

The third factor has high loadings for RM and LSTAT and MEDV, these variables only weakly onto the other factors. The third factor is a ***socioeconomic factor*** *with higher values of this factor associated with more expensive, larger houses and fewer residents in the low socioeconomic category.*

The fourth factor has high loadings for PTRATIO. The variable ZN also loads heavily onto the fourth factor, but not as much as it does for the fourth factor. The fourth factor is possibly a ***education factor*** *with higher values of this factor associated with less investment in education (i.e. too many students per teacher).*

Factor Scores

1. Estimate the factor scores using Bartlett's method when there is no rotation.

```
Boston%>%
  column_to_rownames('Town')%>%
  factanal(factors = 4,rotation = 'none',scores = 'Bartlett')->fa
```

2. Estimate the correlation matrix of the factors.

```
cor(fa$scores)
```

```
##                Factor1       Factor2       Factor3       Factor4
## Factor1  1.000000e+00  2.475727e-16  2.327100e-16 -9.611535e-16
## Factor2  2.475727e-16  1.000000e+00 -4.600106e-16  6.173985e-16
## Factor3  2.327100e-16 -4.600106e-16  1.000000e+00  3.159798e-16
## Factor4 -9.611535e-16  6.173985e-16  3.159798e-16  1.000000e+00
```

3. Repeat the previous 2 questions using the promax rotation

```
Boston%>%
  column_to_rownames('Town')%>%
  factanal(factors = 4,rotation = 'promax',scores = 'Bartlett')->fa_p
cor(fa_p$scores)
```

```
##            Factor1    Factor2    Factor3    Factor4
## Factor1  1.0000000  0.5308492 -0.5220026  0.1532655
## Factor2  0.5308492  1.0000000 -0.4866319  0.2649672
## Factor3 -0.5220026 -0.4866319  1.0000000 -0.2319996
## Factor4  0.1532655  0.2649672 -0.2319996  1.0000000
```

4. Are these answers what you expect? Why or why not?

These answers are as expected. The unrotated factor scores have a correlation matrix that is almost an identity matrix. This implies the factors are uncorrelated. The promax rotation induces correlation in the factors.

PCA v Factor Analysis

How does factor analysis (FA) differ from PCA?

PCA assumes no model (or structure). The objective is simply to identify linear combinations of variables that explain the maximum amount of variance. The principal component are uncorrelated with each other and are not necessarily interpretable. PCA can at most be seen as a crude starting point for EFA. Factor analysis is built around a statistical model. There is a component of variance that is idiosyncratic to each variable and is unexplained by the other variables. Although uncorrelated factors are often assumed this assumption can be relaxed. Factor analysis is much more concerned with interpretation of factors.