HDDA Tutorial: Correspondence Analysis : Solutions

Department of Econometrics and Business Statistics, Monash University

Tutorial 11

Concepts

1. What is the idea behind Correspondence Analysis?

When there are several categories it is difficult to identify patterns and/or associations. Correspondence Analysis tries to reduce the dimension to create a visualisation that can be readily analysed by a researcher. It is somwhat analogous to Principal Components analysis on categorical data. The aim is to find two latent factors that summarise most of the information in the cross tab without much loss of accuracy. Correspondence analysis also allows for a symmetric normalization. This allows comparison of row categories with column categories on the same plot. The distances between the points are on a common scale the closer the points, the higher the association. We cannot test if the individual associations are statistically significant. CA is an exploratory approach.

2. What does inertia measure?

If the observed probabilities were equal to the expected probabilities, inertia would be zero. Inertia therefore, measures the strength of the relationship between the variables in the cross tab. One way to think about it is that inertia measures the amount of information contained in the crosstab. Correspondence Analysis approximates a large dimensional crosstab in (typically) 2 dimensions. In Correspondence Analysis we find a small number of factor scores that explain a large proportion of total inertia. The calculated χ^2 value is inertia times the sample size. So inertia can also be thought of as a measure of dependence within the data.

3. Identify the advantages and disadvantages of correspondence analysis.

Advantages: Simple, can handle nominal data and requires no assumptions.

Disadvantages: Difficult with 3 or more variables; can lose information; cannot look at the effect continuous variables may have on nonmetric variables

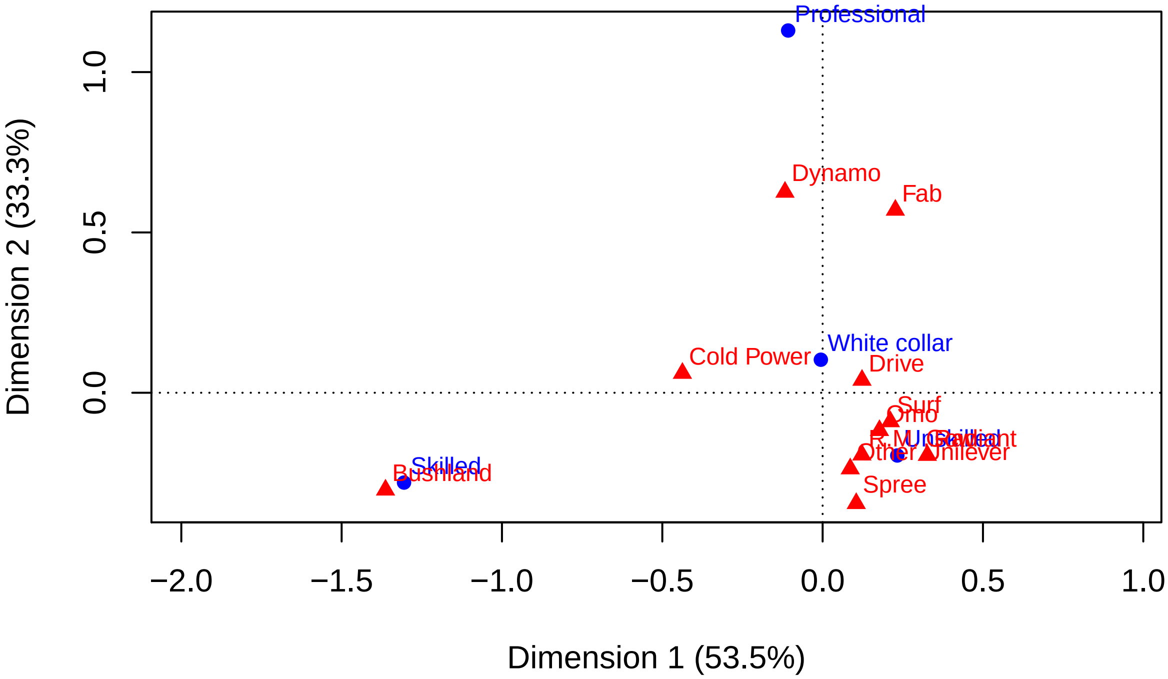
Application

The file Laundry.RData (on the Moodle site) contains data on laundry purchases. Although data are available on many variables, an advertising firm is specifically interested in identifying whether people with particular occupations tend to buy from particular manufacturers.

1. Perform correspondence analysis on these variables.

```r
#Load package
library(dplyr)
library(ca)
#Read Data
Laundry<-readRDS('Laundry.rds')

table(Laundry$OCCUPTN,Laundry$BRAND)%>%
  ca%>%
  plot()
```



2. Name two brands that are similar to one another.

Surf and Omo are similar in the sense that they they have a similar profile of customer occupations.

3. Name a manufacturer associated with skilled occupations.

Bushland is associated with skilled occupations.

4. Name an occupation associated with the manufacturer Unilever.

Unskilled occupations are most closely associated with Unilever.

5. How much inertia is explained by the second dimension on its own.

23.9% of the inertia is explained by the second dimension (this is given in the plot and the summary table in R).