Benchmarks Online

Skip Navigation Links

Page One

Campus Computing News

Helpdesk FYI

Today's Cartoon

RSS Matters

The Network Connection

Link of the Month

WWW@UNT.EDU

Short Courses

IRC News

Staff Activities

Subscribe to Benchmarks *Online*

Research and Statistical Support
University of North Texas

RSS Matters

Ade4TkGUI - A GUI for Multivariate Analysis and Graphical Display in R

Link to the last RSS article here: Tinn-R: A Convenient Script Editor for R on the Win32 Platform - Ed.

By Dr Rich Herrington, ACS **Research and Statistical Support Services Consultant**

This month we take a look at some advanced functionality in R that is available from a drop down menu system in R. Ade4 is a package for multivariate analysis and graphical display for the environmental sciences. Much of this package's functionality will be useful for researchers in the social sciences as well. The package and accompanying documentation can be downloaded from the CRAN website - Ade4, Ade4TkGUI. A more complete tutorial on using Ade4 can be found at the website for Ade4. Methodologies in this package that will be of interest to researchers include **(see Table 1.): Principal Component Analysis**; **Centered and Un-centered Correspondence Analysis**; **Multiple Correspondence Analysis**; Fuzzy **Correspondence Analysis**; **Methods to analyze mixtures of quantitative (interval) and** qualitative variables **(ordinal, categorical)**. Additionally Ade4 implements: **Linear Discriminant Analysis; Canonical Correlation Analysis;** and **statistical tests for between group clusters** based on Monte-Carlo and Permutation based techniques. In this article we will look at how to use Ade4 to implement **"Correspondence Analysis" (CA)**. The technique of CA falls under a wider class of multivariate techniques called "ordination methods" (e.g. Principal Component Analysis; Multidimensional Scaling, etc.). These methods order objects on derived continua (subject to some optimization criteria) such that similar objects are nearer one another, and dissimilar objects are further from one another. Graphical depiction of these derived continua allow graphical based clustering of objects (e.g. row objects and column objects in the data table). The field of Psychometrics has contributed greatly to the development of these methodologies (see the journal Psychometrika). CA is based on a matrix decomposition algorithm called Singular Value Decomposition (SVD) and bears the greatest resemblance to a class of lesser known techniques that are more generally known as "optimal scaling" methods - variously known as: Dual scaling; Optimal Scoring; Reciprocal Averaging; Homogeneity Analysis; or Alternating Least Squares Methods (ALS - see the pseudonym Albert Gifi). These methods derive reduced rank representations (e.g. a reduced set of coordinate systems) or lower dimensional components of transformed categorical and ordinal data by an iterative algorithm that transforms the categorical scaling of the variables

into optimally derived numerical scales. These algorithms (and their variants) are, iteratively applied, constrained least squares optimization procedures (i.e. an iterative application of an eigenvalue/eigenvector (e.g. independent modes of variation in the original scores) extraction algorithm subject to certain row and column constraints - for a readable account see William Jacoby, 1999). As a result, nonlinear multivariate associations between sets of variables can be uncovered. These methods will be valuable in situations where survey researchers are interested in data tables where (for example) the relationship of the rows (respondents) to columns (items or survey questions) are of interest, regardless of whether the rows and columns represent nominal or ordinal level measured variables. These SVD based techniques for categorical or ordinal data bear a theoretical resemblance to techniques for interval level data that allow comparisons between respondents (rows) and items (columns) in "reduced subspaces" (coordinate system), the so called "latent variable" models (for a readable account see Bollen). For example, Factor Analyses (FA) are applied to interval measured data where the measured data (or manifest data) are an indicator of an unobserved, latent, continuous variable. Item Response Theory Analyses (IRT) are applied to ordinal dichotomous, or polytomous ordinal, measured data that are an indicator of an unobserved, latent, continuous variable. Latent Class Analysis (LCA) is applied to data that is nominal or categorical in composition, assuming an underlying latent category for observed responses. What these methods all have in common are that they derive a reduced set of factors, components, or categories for observed or latent scores where the relationship between rows and columns are of interest (e.g. respondent by category; respondent by item; or category by item; item by item; etc.). While the methods in **Ade4** (e.g. **correspondence analysis**) are primarily descriptive and are not model-based (e.g. like Maximum Likelihood Factor Analysis) and do not involve the estimation of sampling variability or interval estimation for parameters, there are some nonparametric based statistical tests available (i.e. resampling or permutation based tests). Additionally, the eigenvalue/eigenvector extraction procedures (i.e. SVD algorithm) and the subsequent common scaling of the coordinate system, do allow researchers to explore respondent and item similarity in a highly useful exploratory graphical procedure (e.g. biplots) of item and respondent similarity. As a set of exploratory methods, these techniques are indispensable for reducing the complexity of multivariate data so that interrelationships amongst sets of variables may be uncovered (respondents, items, and other important covariates). My intention in this article is to demonstrate the steps necessary to produce an analysis with **Ade4** and **Ade4TkGUI** and to demonstrate how ordination based methods can be useful for survey research.

Table I. From R news, http://cran.r-project.org/doc/Rnews/Rnews_2004-1.pdf (page 5)

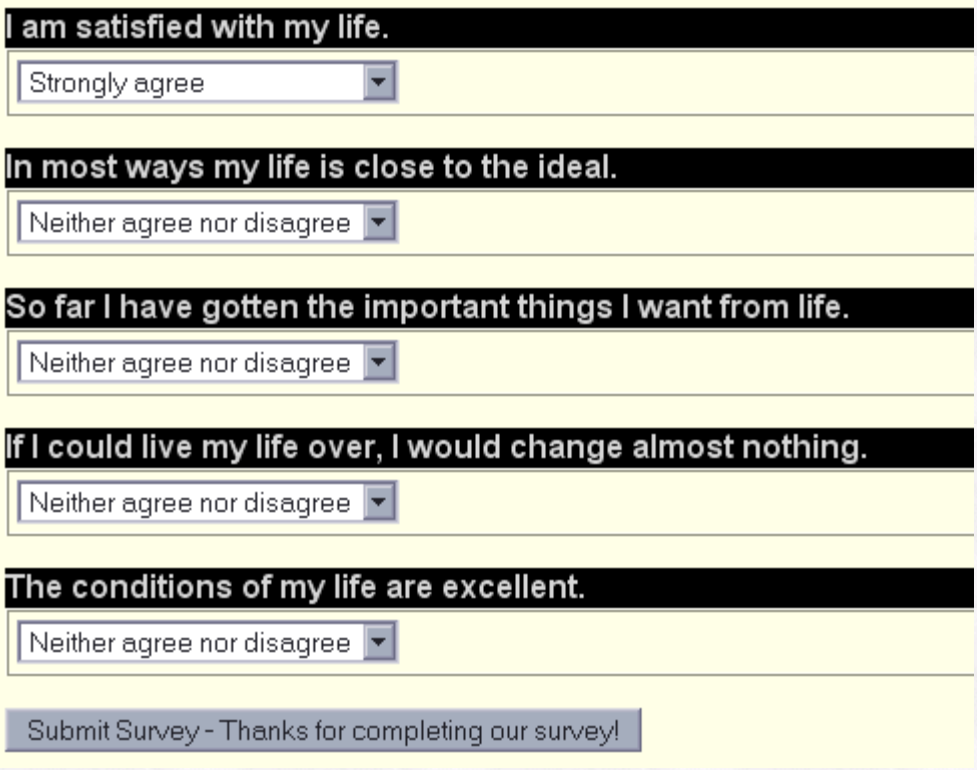
Functions	Analyses	Notes
dudi.pca	principal component	1
dudi.coa	correspondence	2
dudi.acm	multiple correspondence	3
dudi.fca	fuzzy correspondence	4
dudi.mix	analysis of a mixture of numeric and factors	5
dudi.nsc	non symmetric correspondence	6
dudi.dec	decentered correspondence	7

The dudi functions. 1: Principal component analysis, same as prcomp/princomp. 2: Correspondence analysis Greenacre (1984). 3: Multiple correspondence analysis Tenenhaus and Young (1985). 4: Fuzzy correspondence analysis Chevenet et al. (1994). 5: Analysis of a mixture of numeric variables and factors Hill and Smith (1976), Kiers (1994). 6: Non symmetric correspondence analysis Kroonenberg and Lombardo (1999). 7: Decentered correspondence analysis Dolédec et al. (1995).

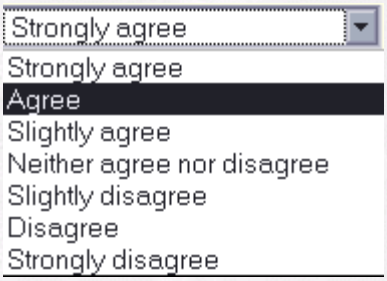
As a working example, we will use the five item survey tool - Satisfaction With Life Survey. Additional information can be found at: http://www.tbims.org/combi/swls/index.html. These data were collected as part of on-going classroom demonstrations here at UNT in both undergraduate and graduate course work.

Correspondence Analysis with the "Satisfaction With Life Scale" (SWLS)

The following are screen shots of the list-drop-down boxes that were used in the SWLS survey. The UNT Zope Survey server was used to collect the responses to the SWLS:



Responses are collected on a 7 point ordinal (arguably an interval) scale; the anchors to the points on the scale are: **Strongly agree, Agree, Slightly agree, Neither agree nor disagree, Slightly disagree, Disagree, and Strongly disagree**:



Some demographic data were collected, however we'll not look at that data here, we'll only examine the responses to the 5 items Q1-Q5. The data are available for download at: http://www.unt.edu/rss/SWLS.questions.txt. Last month we discussed using the Tinn-R editor as a script editor and pager for the R environment on the Windows platform. Below is a screen shot in Tinn-R of the R commands used to: **1) load needed packages; 2) download the data from the URL given; 3) export the survey data to a delimited text file such that the data can be read in by the ade4 GUI interface; and 4) display the data in a window for examination**. Additionally, the **ade4TkGUI** menu is displayed as well. Below is a screen-shot Tinn-R with the necessary R commands. Additionally a screenshot of 16 out of 174 responses (as an example) to the SWLS are displayed below the Tinn-R screenshot.

R Commands needed to load packages; download data; and start the ade4 GUI:

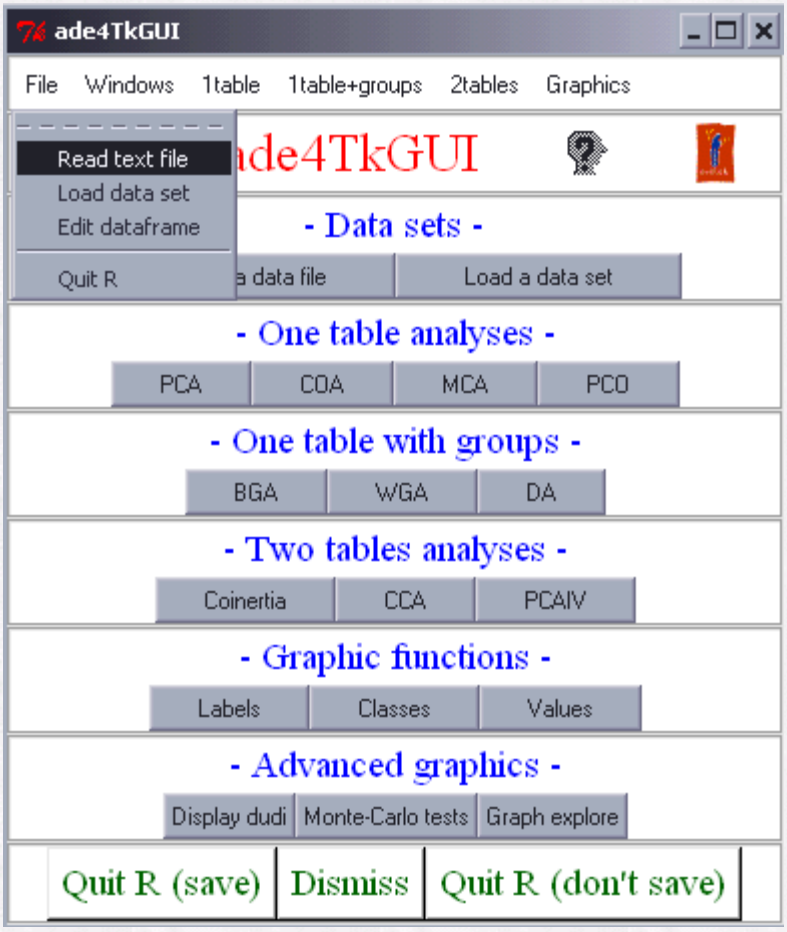
```
# Load libraries
library(relimp)
library(ade4TkGUI)

# Read data
SWLS.dat<-read.delim("http://www.unt.edu/rss/SWLS.questions.txt")

# Write out as delimited text
write.table(SWLS.dat,
            "C:/SWLS.questions.dat.txt",
             sep="\t", col.names=TRUE, row.names=FALSE,
             quote=TRUE, na="NA")

# Display spreadsheet
showData(SWLS.data)

# Load multivariate library
ade4TkGUI()
```

16 out of 174 responses are displayed:

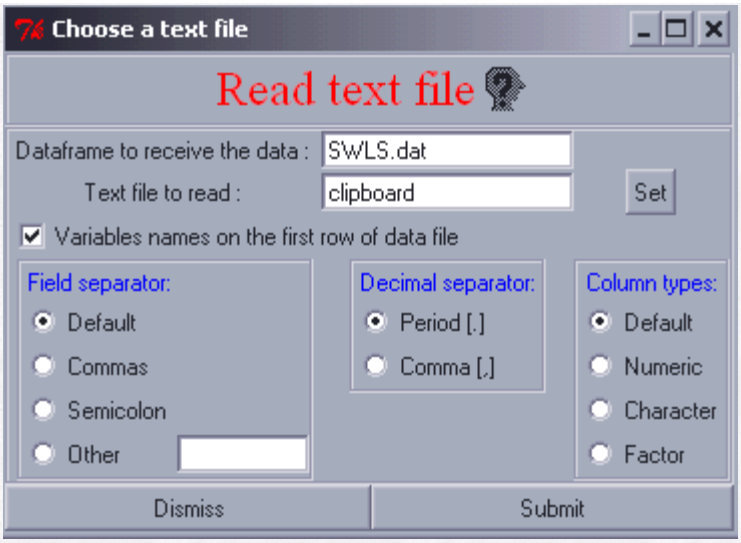
SWLS.data

	Q1	Q2	Q3	Q4	Q5
1	2	3	2	3	6
2	1	1	2	2	2
3	1	1	1	2	5
4	4	4	6	2	3
5	2	3	4	2	5
6	2	2	2	2	6
7	6	4	5	5	5
8	2	1	2	1	1
9	5	5	6	6	6
10	5	5	5	3	4
11	1	1	1	1	1
12	6	5	6	6	7
13	4	4	4	4	4
14	3	4	5	3	2
15	2	3	3	2	6
16	2	3	4	3	2

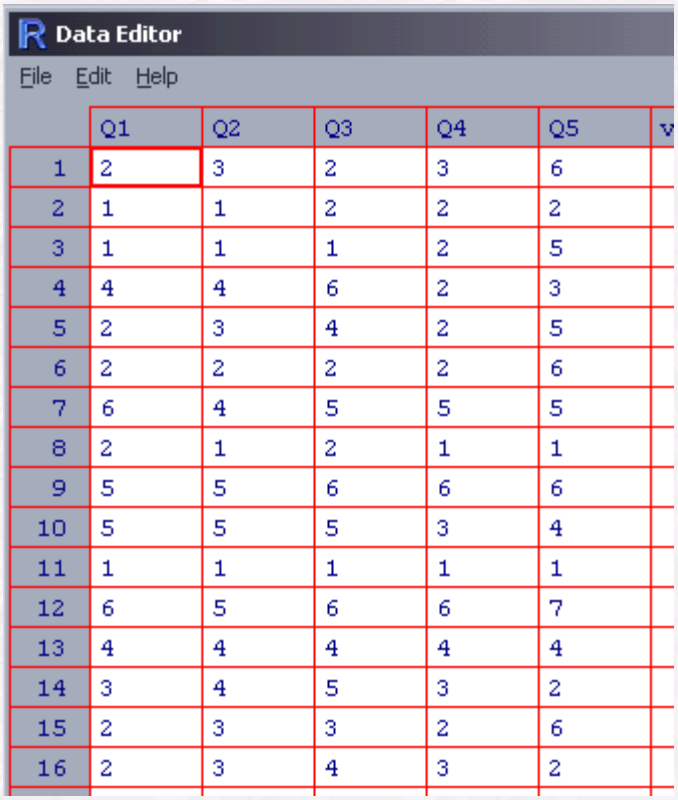
The Initial ade4 GUI:



To begin, we need to read the data into the R working environment: On the Ade4 main menu bar, select: **File - Read Text Data** which will bring up the following window:



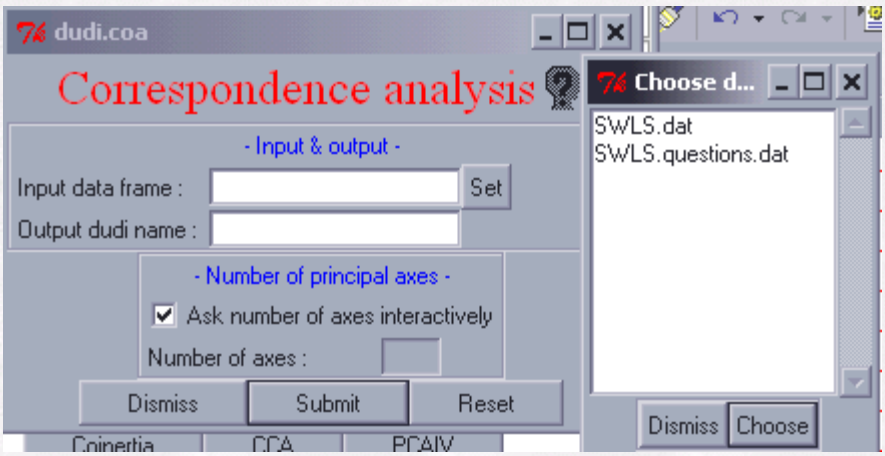
The defaults will work for the tab-delimited file SWLS data file (with variables on row one). Give the dataset a name of **SWLS.dat.** This will be the data set name for the workspace that we are using in R. Select the **"Set"** button and **browse to the location of the SWLS text file** (for this example the data was exported to **"c:\" drive**). Once the data is selected, a data editor appears. Close this window after a visual inspection - make sure the data loaded properly. When you are finished inspecting the data **close the R data editor by clicking the "x" in the upper right corner of the data editor**. For example:



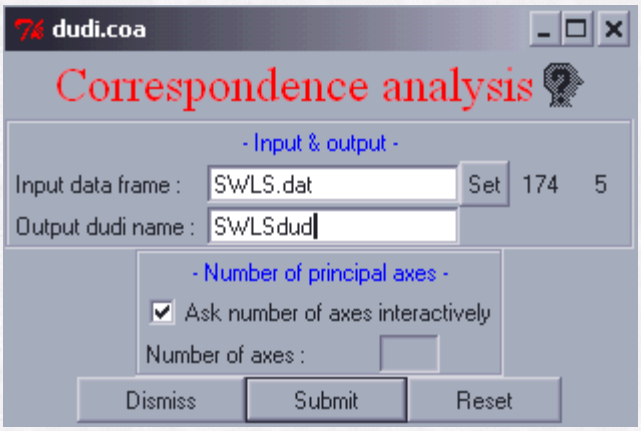
	Q1	Q2	Q3	Q4	Q5	v
1	2	3	2	3	6	
2	1	1	2	2	2	
3	1	1	1	2	5	
4	4	4	6	2	3	
5	2	3	4	2	5	
6	2	2	2	2	6	
7	6	4	5	5	5	
8	2	1	2	1	1	
9	5	5	6	6	6	
10	5	5	5	3	4	
11	1	1	1	1	1	
12	6	5	6	6	7	
13	4	4	4	4	4	
14	3	4	5	3	2	
15	2	3	3	2	6	
16	2	3	4	3	2	

The "read-text" dialog box should disappear and the display returns to the initial Ade4 dialog box. In the "One table analyses" panel, **Select the "COA"** button for the **"correspondence analysis"** method:

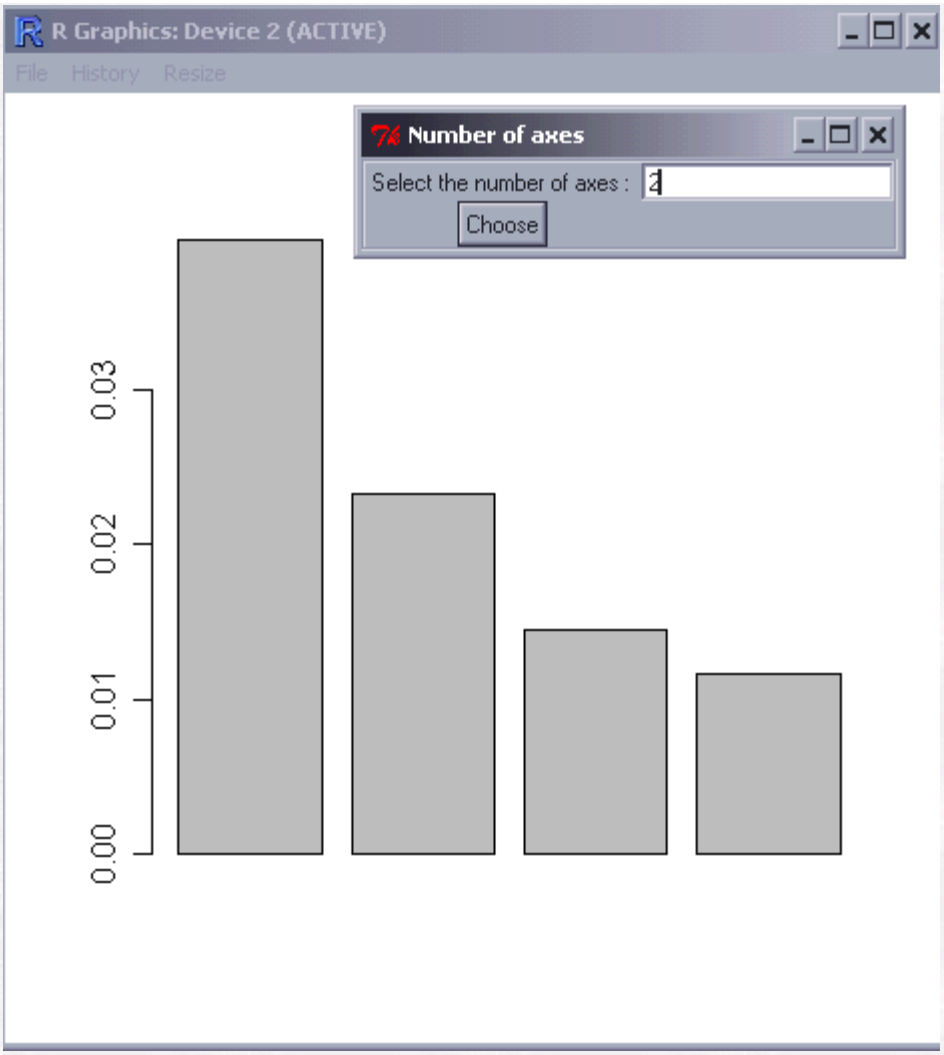
On the following dialog box, click **"Set"** and fill the **"Input data frame"** field by selecting the **SWLS.dat** entry in the next popup window. Then click **"Choose"** and then **"Submit":**



We see:

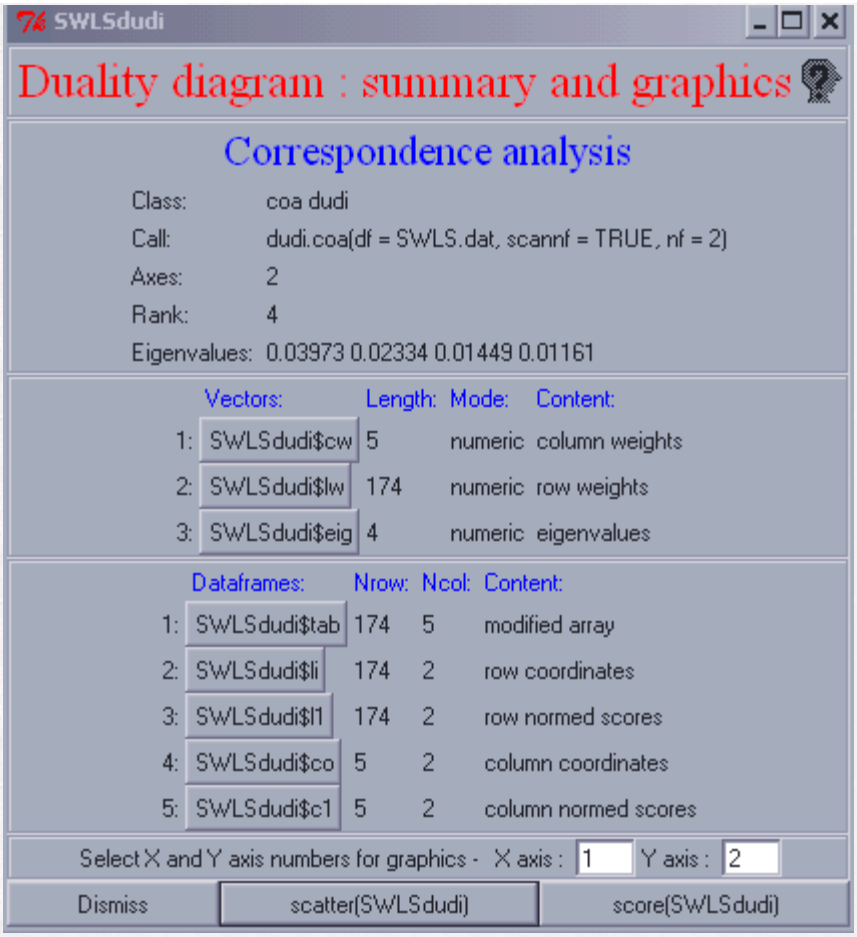


The dialog box shows that 174 rows with 5 columns have been selected. Give an **"output dudi name"** of **"SWLSdudi"** and click **"Submit"**. The following windows are generated:

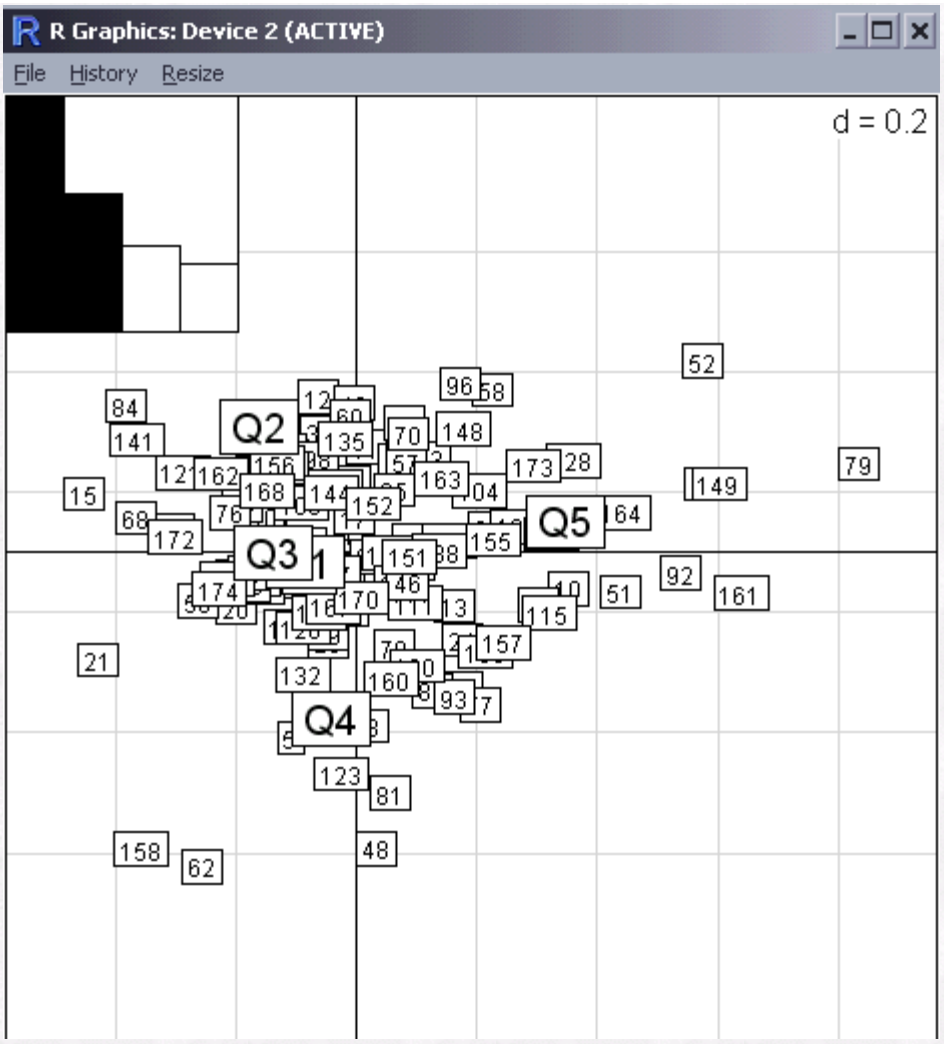


Select **"2"** axis for the display and click **"Choose"**. **The bar chart that is displayed is a representation of the number of modes of independent variation accounted for in the original scores, where the heights display the relative amounts of variation accounted for in the original scores**. By selecting, "2", we are choosing to plot row objects and column objects in a coordinate system scaled such that the first two independent modes of variation are represented by the **(x, y)** axes of the coordinate system. By having the axes scaled in a common metric, row and column objects can be compared in terms of distances.

The following dialog box appears:



This dialog box gives us information regarding the eigenvalue/eigenvector extraction and row and column scalings that Ade4 performed ("Vectors" panel). In the Vectors panel we see: column weights, row weights, and eigenvalues. Four eigenvalues were extracted with values of: .039, .023, .014, and .011 (note: since there are 5 items only 5 eigenvalues could be extracted; only the set larger than .01 are displayed). Clicking the **"scatter(SWLSdudi)"** button produces an **(x, y)** plot of the **case and item scores** (labeled with row and column number IDs) where the scaling for the **(x, y)** coordinates are equivalent. The upper left panel of the window displays a shaded bar graph of the two eigenvalues that are producing the **scores** and **scalings** of the coordinate system **(x, y), which are the two largest modes of variation in the row (respondents) and column (items) objects**:



Initial Interpretations of the Correspondence Analysis

The pair of axes **(x, y)** represent independent coordinates or uncorrelated "components of variation", or "units of information" for the row and column objects. The units are scaled in a common metric for both **x** and **y** axes. Which, as a set of **(x ,y)** pairs, describe each of the row objects (n=174 respondents), and column objects (5 items). For a set of perfectly homogeneous items, we would expect the items to cluster fairly close to one another on both axes, with most of the clustering occurring along one axis. Since these items were semantically constructed to elicit the self report of a theoretically defined variable called **"Life Satisfaction"**, **we expect a single or unidimensional construct to emerge across individuals such that items look similar in terms of the (x, y) coordinates**. That is, we expect all of the information in the original set of elicited responses to be contained in the transformed values **(x, y)** with most of the information contained in either **x** or **y**. We would expect that the **(x, y)** values would be close for the items that are more homogeneous (i.e. responded to similarly by respondents). Additionally, individual respondents who are close in their **(x, y)** pairs would be considered to be more similar in their response patterns across the set of items as opposed to individuals whose **(x, y)** values differ substantially. Moreover, **the closeness of individuals AND items on (x, y) scores would allow a researcher to cluster "similar individuals" on clusters of "similar items"**. In our survey of 5 items, we see that items Q1, Q3 and Q5 are more similar to one another on the **x coordinate**. And, Q1, Q2, Q3, and Q4 are more similar to one another on the **y coordinate**. **Item Q5 can be seen as standing apart from the other four items (even more so than Q2)**. Also notice

that **the bulk of the respondents fall near items Q1-Q4. Item Q5 might be better reworded or be discarded entirely**. One way of thinking of this situation is to **see that there are 3 seperate TYPES of questions: 1) Low (x) values - [Q1,Q2,Q3,Q4]**; and **2) Low (y) values - [Q1, Q3, Q5];** and **3) Low (x, y) values - [Q1, Q3]**. These patterns of variation would account for three of the largest eigenvalues (e.g. indpendent modes of variation in the original scores). **Perhaps the smallest eigenvalue is accounted for by item Q2** since it resides some distance from both the **x** and **y** axes to some extent. Notice that **respondents 21, 48, 52, 62,79, 158, & 161** reside a substantial distance away from the bulk of the other respondents.

Conclusion

In summary, our conclusion is that this Correspondence Analysis has helped reveal a potentially informative source of heterogeneity in the set of items and respondents (rows and columns). The original presupposition of a **unidimensional construct underlying these items does not seem to hold**, at least upon a graphical inspection (and is supported by multiply large eigenvalues). Our **next step might be to look for subgroups of individuals** that account for the heterogeneity that we see in respondents responses on certain items (e.g Q5). Additionally, **we might try to clarify the wording in the survey items to better communicate the semantic content** that we are hoping will elicit correlates of the construct "Life Satisfaction" in our respondents responses.

References

Statnotes: Topics in Multivariate Analysis, by G. David Garson - Correspondence Analysis Section

Multivariate Statistics: Concepts, Models, and Applications, by David Stockburger - Linear Transformations Section

Statistics With R, by Vincent Zoonekynd - Factor Analysis Section

Special Announcements: RSS will be maintaining a blog devoted to research and statistics related news - RSS-Blogs; Additionally, RSS will be maintaining a Zope/Plone website devoted organizing communities and resources involved in survey research - RSS-Surveys.

Return to top