Correspondence Analysis

Graphical procedure for representing associations in a contingency table

Introduction

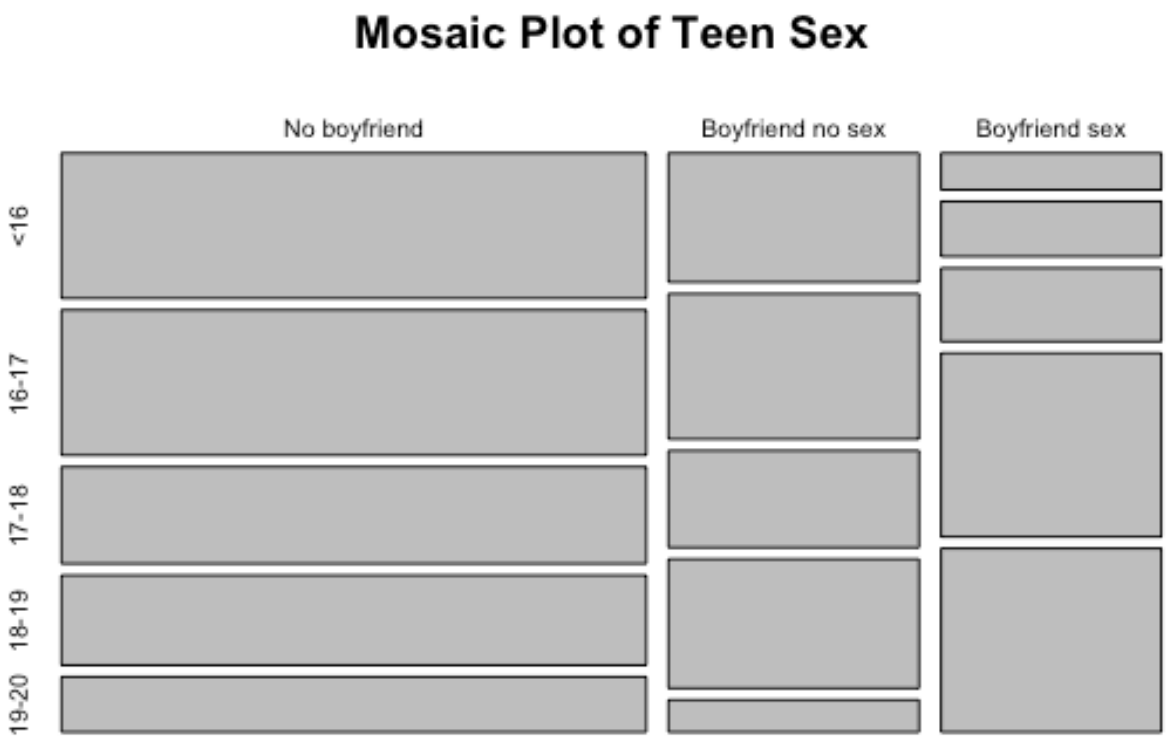
- Let the two-way contingency table have I rows and J columns.
- Then the correspondence analysis plot contains two sets of points: I points corresponding to the rows and J points corresponding to the columns.
- Row points that are close together indicate rows that have similar profiles (aka conditional distributions).
- Column points that are close together indicate columns that have similar profiles.
- Row points that are close to column points indicated combinations that occur more often than what is expected under independence.

Example: 4.6.1 on p. 130 of Everitt & Hothorn

	<16	16-17	17-18	18-19	19-20
No boyfriend	21	21	14	13	8
Boyfriend no sex	8	9	6	8	2
Boyfriend sex	2	3	4	10	10

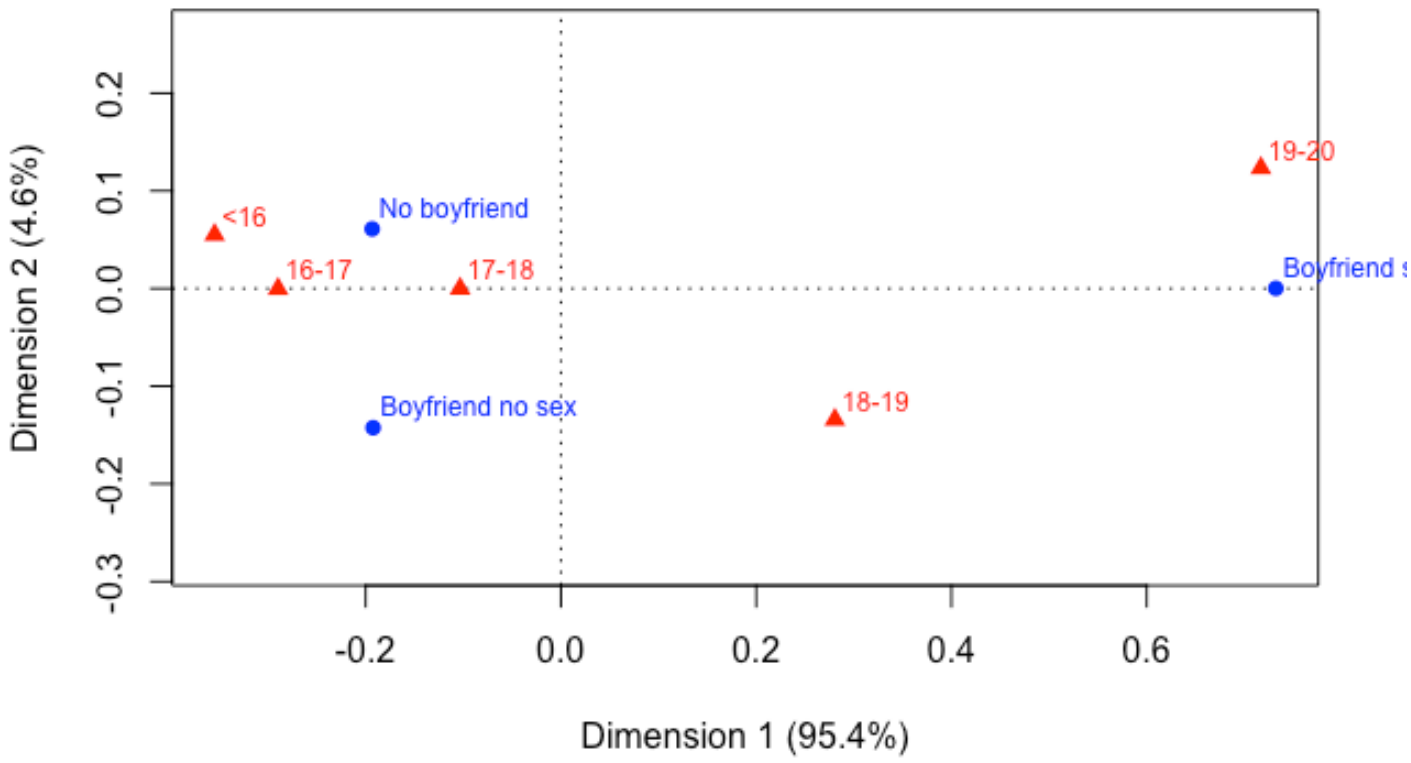
- $I = 3$ types of relationships
- $J = 5$ age groups
- If we divide the frequencies by the row or column totals we obtain the profiles aka conditional distributions
- If the two variables are independent the profiles should be very similar.

Mosaic Plot of Profiles



In general, profiles look different.

Correspondence Analysis Plot



- Age groups "< 16" and "16-17" have similar profiles
- "19-20" and "Boyfriend sex" are close together and away from the rest of the points

Algebra

Let $\boldsymbol{X}$ be a $I \times J$ two-way contingency table with observed counts x_{ij} and $I > J$. Define n to be the total count of all frequencies. Construct the the proportions matrix $\boldsymbol{P} = \{p_{ij}\}$ as follows:

$$p_{ij} = \frac{x_{ij}}{n}, i = 1, \ldots, I, j = 1, \ldots, J \text{ or } \boldsymbol{P} = \frac{1}{n}\boldsymbol{X}$$

Next define the vectors of row and column totals $\boldsymbol{r}$ and $\boldsymbol{c}$:

$$r_i = \sum_{j=1}^{J} p_{ij} = \sum_{j=1}^{J} \frac{x_{ij}}{n}, i = 1, \ldots, I \text{ or } \boldsymbol{r} = \boldsymbol{P}\boldsymbol{1}_J$$

$$c_j = \sum_{i=1}^{I} p_{ij} = \sum_{i=1}^{I} \frac{x_{ij}}{n}, j = 1, \ldots, J \text{ or } \boldsymbol{c} = \boldsymbol{P}'\boldsymbol{1}_I$$

$\boldsymbol{D}_r = \text{diag}(r_1, \ldots, r_I)$ and $\boldsymbol{D}c = \text{diag}(c_1, \ldots, c_J)$

Algebra (continued)

Define the square root matrices for scaling:

$$\boldsymbol{D}_r^{1/2} = \text{diag}(\sqrt{r_1}, \ldots, \sqrt{r_I}), \qquad \boldsymbol{D}_r^{-1/2} = \text{diag}\left(\frac{1}{\sqrt{r_1}}, \ldots, \frac{1}{\sqrt{r_I}}\right)$$

$$\boldsymbol{D}_c^{1/2} = \text{diag}\left(\sqrt{c_1}, \ldots, \sqrt{c_J}\right), \qquad \boldsymbol{D}_c^{-1/2} = \text{diag}\left(\frac{1}{\sqrt{c_1}}, \ldots, \frac{1}{\sqrt{c_J}}\right)$$

Goal: Find a reduced rank matrix $\widehat{\boldsymbol{P}} = \{\hat{p}_{ij}\}$ which minimizes

$$\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(p_{ij} - \hat{p}_{ij}\right)^2}{r_i c_j} = \text{tr}\left[\left(\boldsymbol{D}_r^{-1/2}(P - \hat{P})\boldsymbol{D}_c^{-1/2}\right)\left(\boldsymbol{D}_r^{-1/2}(P - \hat{P})\boldsymbol{D}_c^{-1/2}\right)'\right]$$

Result: The reduced rank s approximation is given by

$$\widehat{\boldsymbol{P}} = \sum_{k=1}^{s} \lambda_k \left(\boldsymbol{D}_r^{1/2}\boldsymbol{u}_k\right)\left(\boldsymbol{D}_c^{1/2}\boldsymbol{v}_k\right)'$$

where $\boldsymbol{\lambda}_k$, $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$ are the values and vectors in the SVD of $\boldsymbol{D}_r^{-1/2}\boldsymbol{P}\boldsymbol{D}_c^{-1/2}$

Symmetric Map

It is usual in correspondence analysis to plot the first two columns of

$$\boldsymbol{F} = \boldsymbol{D}_r^{-1}\left(\boldsymbol{D}_r^{1/2}\boldsymbol{U}\right)\boldsymbol{\Lambda}$$

and

$$\boldsymbol{G} = \boldsymbol{D}_c^{-1}\left(\boldsymbol{D}_c^{1/2}\boldsymbol{V}\right)\boldsymbol{\Lambda}$$

It is called a *symmetric map* since the points representing the rows and columns have the same normalization or scaling. That is, the geometry for the row points is identical to the geometry of the column points.

Inertia

The total inertia is a measure of the variation in the count data and is defined as follows:

$$\text{tr}\left[\left(\boldsymbol{D}_r^{-1/2}(\boldsymbol{P}-\boldsymbol{r}\boldsymbol{c}')\boldsymbol{D}_c^{-1/2}\right)\left(\boldsymbol{D}_r^{-1/2}(\boldsymbol{P}-\boldsymbol{r}\boldsymbol{c}')\boldsymbol{D}_c^{-1/2}\right)'\right]=\sum_{k=1}^{J-1}\lambda_k^2$$

The inertia associated with the best reduced rank $s < J$ approximation is

$$\sum_{k=1}^{s}\lambda_k^2$$

This is typically displayed along the coordinate axis.

Example: Household tasks

	Wife	Alternating	Husband	Jointly
Laundry	156	14	2	4
Main_meal	124	20	5	4
Dinner	77	11	7	13
Breakfeast	82	36	15	7
Tidying	53	11	1	57
Dishes	32	24	4	53
Shopping	33	23	9	55
Official	12	46	23	15
Driving	10	51	75	3
Finances	13	13	21	66
Insurance	8	1	53	77
Repairs	0	3	160	2
Holidays	0	1	6	153

Plot of results