

Modelling Categorical Data by Interactive Mosaic Plots and Tables

Martin Theus, University of Augsburg, Germany
Adalbert Wilhelm, University of Augsburg, Germany

ABSTRACT: Categorical data arise in many investigations. Usually, such data will be displayed in contingency tables and analysed using loglinear models and χ^2 -type-statistics. However, the visualisation of more than three variables is very poor in contingency tables. In the modelling phase, often, imputations based upon prior knowledge are made to select relevant variables and interactions instead of looking at the raw data. Interactive Statistical Graphics and Tables can reach beyond those classical limitations.

KEYWORDS: Loglinear Models; Mosaic Plots; Stepwise graphical selection; Interactive contingency tables; Table alternations

1 Introduction

Mosaic Plots (Hartigan & Kleiner, 1981) as well as interactive contingency tables offer the possibility of getting both a graphical and numerical insight into the data. Interacting with Mosaic Plots allows fast and efficient alternation of the multivariate views. Embedding the residuals of a loglinear model into the plots leads to new modelling techniques. The techniques shown here extend Friendly's (Friendly, 1994) approach not only by means of interactivity.

Interactive contingency tables provide the user with the facility of easily switching between low-dimensional views of the multivariate data. Using raw data as well as local and global summands of the power-divergence-statistics (Read & Cressie, 1988) they allow the detection of unusual cells that have a strong impact on the resulting statistics. Interpretation of significant effects is straightforward by combining categories and collapsing tables.

Graphical and table based techniques are implemented as software tools MANET (Unwin, (1996) for mosaic plots) and TURNER (for interactive tables). Furthermore MANET and TURNER can communicate to exchange models as well as graphs.

2 The Data

The most common example used to illustrate the modelling of multivariate categorical data is the detergent dataset first described in Ries & Smith (1963).

The survey includes four variables. Three of them: watersoftness (soft, medium, hard)(1), temperature (low, high)(2) and previous-use (brand X, brand M)(3) are design variables and preference (brand X, brand M)(4) is the response variable. The data are further discussed in e.g. Cox & Snell (1989), Venables & Ripley (1994). None of the articles shows the raw data graphically, only few of them give a plot of the observed against the predicted values, a plot that is, usually, not suitable for any interpretation purpose. The obtained models differ a lot from author to author and, usually, no attempt is made to discriminate between concurrent models.

3 Graphical Model Selection based on Mosaic Plots

Modelling based on Mosaic Plots can be done in two directions. Whereas the graphical backward selection starts with a model with all interactions included, the graphical forward selection starts with mutual independence and adds interactions step by step using the superimposed residual information.

3.1 Graphical Backward Selection

The interaction structure between two or more variables can be seen in Mosaic Plots easily. Figure 1 shows an example of an interaction and independence between two variables. Independence inside Mosaic Plots can be seen by straight

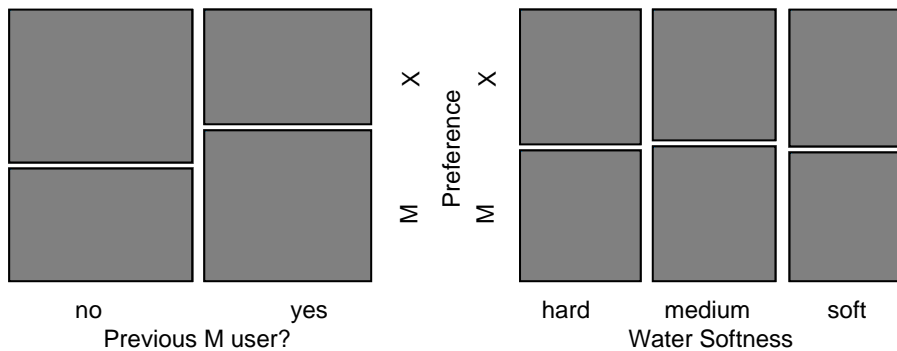


FIGURE 1. Examples for interaction (left) and independence (right)

lines dividing the categories of all variables, c.f. figure 1 right. Recognizing independence and interaction structures (partial independence, conditional independence, no three-way- interaction) in Mosaic Plots even for more than two dimensions can be generalised easily.

Starting with the model including all interactions: $u + u_1 + u_2 + u_3 + u_4 + u_{12} + u_{13} + u_{14} + u_{23} + u_{24} + u_{34} + u_{123} + u_{124} + u_{134} + u_{234} + u_{1234}$ we can deselect all twoway interactions, which are not present. To obtain strictly hierarchical models, all higher interactions including the two indices must be deselected, too. In the detergent example the interaction of water softness vs. preference is

not present. Deselecting u_{14} yields: $u + u_1 + u_2 + u_3 + u_4 + u_{12} + u_{13} + u_{23} + u_{24} + u_{34} + u_{123} + u_{234}$. Proceeding like this yields two more independencies: previous use of M vs. temperature (u_{23}), and previous use of M vs. water softness (u_{13}) and the model: $u + u_1 + u_2 + u_3 + u_4 + u_{12} + u_{24} + u_{34}$. All remaining interactions can be seen in the Mosaic Plots very clearly.

3.2 Graphical Forward Selection

Graphical forward selection starts with the model of mutual independence. By superimposing the residuals in the mosaic plot of the expected values by coloured or shaded (in black & white printing) highlighting, the interaction structure can be displayed. Note the different approach to Friendly (1994), by displaying the mosaic for the model, but not for the raw data. Comparing figure

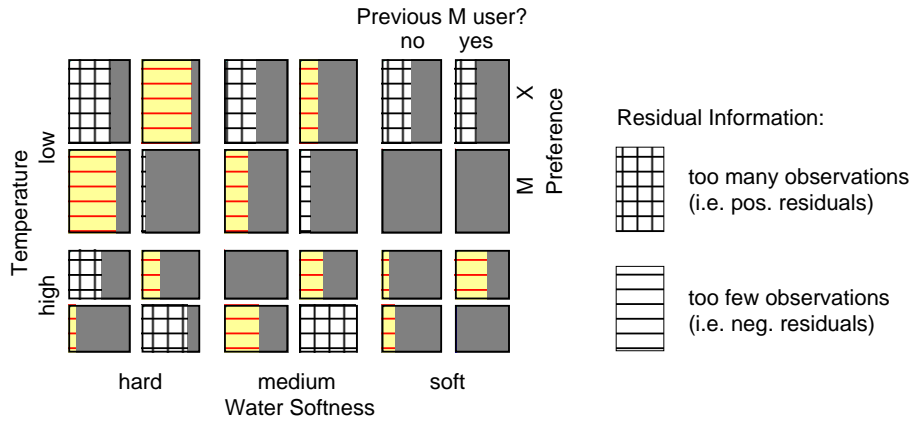


FIGURE 2. Mosaic Plot for a simple model, residuals included

2 shows the overcrossing structure of both positive and negative residuals in the third (previous use of M) and the fourth (Preference) variable. This indicates a high interaction between these two variables. Inclusion of this interaction can be done by simply selecting this lower dimensional pair of variables, and yields the model: $u + u_1 + u_2 + u_3 + u_4 + u_{34}$. Proceeding, i.e. looking at the residual of the new model, suggests two more interactions, yielding: $u + u_1 + u_2 + u_3 + u_4 + u_{12} + u_{24} + u_{34}$ which is the same model as in the backward selection technique. The interpretation of the '+' or '-' shading corresponds to the pressure model given by Friendly (1995).

4 On the use of Interactive contingency tables

Interactive contingency tables are based on projections or slices. Projection means that a two-way contingency table is built and data is collapsed over the other variables. Using projections one has to be aware of Simpson's paradox. Slicing means building one single two-way contingency table for each combina-

tion of categories in the non-displayed variables. Information can be obtained from the raw data, the (local) χ^2 -summands and the global χ^2 -summands. Interactive contingency tables are characterised by easy switching between these different kinds of measures and by the ability of flexibly combining categories.

4.1 Modelling

Similar to Mosaic Plots Interactive contingency tables can be used for forward and backward selection. In the backward selection mode sliced and projected contingency tables can be used to determine independency or conditional independency of variables. The local χ^2 -summands are more appropriate for this method. The forward selection mode can be performed with both local and global χ^2 -summands. Starting with the naive model those interactions are included that correspond to cells with large absolute value of the χ^2 -summands.

4.2 Outliers, Diagnostics & Interpretation

Often contingency tables contain one cell that does not fit the model, whereas the model is adequate for all other cells. Interactive contingency tables makes it easy to ignore some striking cells in fitting the model. Also leverage effects of some cells can be seen from these tables. After significance of some interactions is found, questions of how to interpret the interaction and where it stems from are posed. By being able to delete and combine categories in Interactive contingency tables these tasks are considerably simplified.

References

- Cox, D.R. & Snell, E.J. (1989) *The Analysis of Binary Data*. Second Edition. London: Chapman & Hall.
- Friendly, M. (1994), Mosaic Displays for Multi-Way Contingency Tables, *Journal of the American Statistical Association*, Vol. 89, No. 425.
- Friendly, M. (1995), Conceptual and Visual Models for Categorical Data, *The American Statistician*, Vol. 49, No. 2.
- Hartigan, J.A. & Kleiner, B. (1981), Mosaics for Contingency Tables, *Computer Science and Statistics*, ed. W.F. Eddy, Springer, New York, 268–273
- Read, T.R.C. & Cressie, N.A.C. (1988), *Goodness-of-Fit Statistics for Discrete Multivariate Data*. New York: Springer.
- Ries, P.N. & Smith, H. (1963), The use of chi-square for preference testing in multidimensional problems. *Chemical Engineering Progress*, 59, 39–43.
- Unwin, A. R. (1996), Interactive Graphics for Data Sets with missing values – MANET. *Journal of Computational and Graphical Statistics*, 4, No. 6 .
- Venables, W.N. & Ripley, B.D. (1994) *Modern applied statistics in S-PLUS* Berlin, Heidelberg, New York: Springer.