Using Glim

J. A. Nelder

Introduction

The analyses were done using GLIM (Payne et al., 1983) and log linear models. They are incomplete because of pressure of time and frustrations caused by computing problems. My first aim was to develop a model which showed the main pattern for the initial set of data (France 1989).

The table was split into two components, the diagonal elements and the off-diagonal elements. Two tables are formed: the first is the aggregated switching data, a 2*15 table showing brand*(switching vs. non-switching). The second table contains the original table without the diagonal elements and shows how those who switched allocated their choices among the other brands. (For computing purposes the second table is just the complete table with the diagonal elements weighted out.)

The Aggregated Switching Data

The total number not switching for each brand is given by the diagonal elements of the table. The total number switching to another brand in the table is given by the rest of the column totals. The figures are

	Alfa	BMW	Cit	Fiat	Ford	GM	Lada	Mer	Peu	Ren	Rov	Saab	Seat	VW	Volv
no switching	97	163	1811	526	696	362	68	136	2928	4861	115	10	36	772	78
switching	162	197	1455	613	582	323	103	59	1949	2799	212	9	101	606	86

The percentage switching, ranked in order, is

Ren	Cit	Peu	Alfa	BMW	Mer	VW	Volv	Ford	Fiat	Saab	Lada	GM	Rov	Seat
28	33	40	44	47	48	54	54	55	62	66	66	66	68	89

Note the three French brands come first and also that the three German brands are next to each other. Seat is outstandingly high.

The Detailed Switching Data

The main idea here was to see if the brands could be split into fairly homogeneous subgroups in respect of the switching proportions within them, and then to look at the patterns between groups. The hypothesis of symmetry, which is often entertained about tables like this where the same factor indexes two or more dimensions, is here less than compelling because there seems little reason a priori why the switch from A to B should be the same as from B to A.

The main technique was to fit a log-linear model using the current split and then to examine the extreme residuals (both positive and negative) for the presence of pairs of the form (i,j) and (j,i).

Such pairs would suggest the presence of a further subgroup that should be explored.
Treating all brands as one group we fit an additive model on the log scale with Poisson errors (i.e. a log-linear model) and find a deviance of 705.6 with 181 d.f. This gives a mean deviance of about 4, showing considerable overdispersion. Examination of the residuals shows that the two

extreme positive residuals are (BMW,Mer) & (Mer,BMW), both German makes, and that (VW,BMW), involving the third German make is also large. I extracted the German brands as a subgroup, and refitted the model using separate sets of parameters for the four sections of the table defined by (German, rest) * (German, rest). The deviance is now 381.6 with 154 d.f., a very substantial fall: inspection of the residuals from this fit shows a cluster involving the set (Cit,Ford,GM), i.e. the two U.S. brands and Citroen. If we remove all rows and columns involving (BMW,Mer,VW) then all six residuals from the set (Cit,Ford,GM) appear among the extreme residuals. With two subgroups BMW,Mer,VW) and (Cit,Ford,GM) and a third being the rest, the deviance falls to 262.2 with 129 d.f. I note the presence of (Rov,GM) as giving the largest negative residual. The data value for this point is 3 and the fitted value is 13.7. This point is extreme with many models and suggests that there might be a gross error in recording here. It would be interesting to know if this is so, particularly when the other tables have values nearer the fitted 13.

It is possible to continue the process of looking for further subsets based on the presence of extreme residuals; the problem is that it is uncertain what the baseline deviance is. Almost certainly there will be some inherent heterogeneity leading to overdispersion; the value of about 2 for the men deviance after extracting two subgroups is not excessive.

It would be possible to look at the matter of symmetry by fitting models in which symmetry conditions within each subset are imposed, as compared with the above models, where symmetry is not assumed.

The Analysis of Several Data Sets

The main interest will center on whether the pattern shown in the first is repeated in the others. The first question is whether the subsets identified in the first set are found in the others. If there is substantial, but not complete, agreement between sets, then a compromise grouping should be sought which gives adequate fits over all sets when compared with the best individual fits. If such an overall split into subsets can be found then the second question concerns the consistency of the parameter estimates over the sets using a common grouping. This can be done in the usual way by comparing the fit using common parameter values over all sets with the individual best fits, and inspecting the individual deviations to see if any one set is discrepant from the pattern defined by the others.