

Using SAS® PROC GENMOD for Log-linear Association Models of Ordinal Data
James R. Hershey, Department of Psychology, University of Illinois at Urbana-Champaign

ABSTRACT

This paper illustrates the use of the GENMOD procedure to analyze ordinal data. Models which take into account the ordinal nature of one or both variables in a two-way table are discussed. The SAS DATA step is used to obtain the right structure in the data, and the code for several models fit with PROC GENMOD is shown. An example illustrates one way to extract information about the association between two variables that would be missed with ordinary contingency table analysis.

1 INTRODUCTION

Log-linear models are commonly used for data that are expressed as frequencies. The most widely used display of this type of data is the cross-tabulation of the frequencies by row and column. The variables representing the rows and columns are discrete values. When these discrete values contain ordered categories, traditional log-linear (nominal) and general linear models (e.g., ANOVA and regression) are often not optimal to analyze these data. ANOVA and regression models require inappropriate assumptions about these data, such as being continuous and normally distributed. The traditional log-linear model is an appropriate type of model, but it is not able to detect ordinal association. The log-linear model for a two-way table only tests the no-interaction model vs. the saturated model. This is equivalent to the familiar Pearson or likelihood-ratio tests of independence. However, this test only indicates lack of independence, but does not reveal the nature of the association. Work developed by Goodman (1979) and illustrated in Clogg & Shihadeh (1994) shows how ordinal information may be incorporated into log-linear models. These models may be used for data that have one nominal and one ordinal variable, as well as when both variables are ordinal. The SAS data step can be used to easily create the variables for the appropriate design matrix representation of the model. The GENMOD procedure is used to illustrate how to fit these models.

2 STATISTICAL MODELS

The saturated log-linear model is represented by

$$\log F_{ij} = \lambda + \lambda_i^R + \lambda_j^C + \lambda_{ij}^{RC},$$

which fits the data perfectly and has no degrees of freedom. The λ_i^R variables represent the marginal distribution of the rows, the λ_j^C fit the column marginal distribution, and the λ_{ij}^{RC} are the association parameters. The independence model simply removes the association parameters:

$$\log F_{ij} = \lambda + \lambda_i^R + \lambda_j^C.$$

Thus, the expected values are obtained from the marginal distributions. The degrees of freedom in this case are $(I-1)(J-1)$.

When there are ordinal variables, they must be scored quantitatively. Common methods of scoring are to use integers (i.e., Low, Medium and High become 1, 2 & 3), category midpoints or category midranks. Once the data have been scored, a log-linear model known as the Linear X Linear Interaction model can be fit to the data:

$$\log F_{ij} = \lambda + \lambda_i^R + \lambda_j^C + \phi x_i y_j,$$

where the x_i and y_j are the row and column scores, respectively. The marginals are fit as before, but the association between the row and column variables is represented by one parameter, ϕ . This model is tested as before, but now has one less degree of freedom. The case where integer scoring is used for both variables is known as the Uniform Association model. This is because all local odds ratios will be equal to one value, θ . In the case when independence holds, all local odds ratios will be equal to one.

When there are one nominal and one ordinal variable, a row (column) effects model can be fit. The row (column) is treated as nominal, and the column (row) is treated as a scored ordinal variable. Here we must fit an association parameter for each row (column). The Row Effects model is

$$\log F_{ij} = \lambda + \lambda_i^R + \lambda_j^C + \sum_{i=1}^{I-1} \phi_{R(i)} D_{R(i)} y_j.$$

Here the $D_{R(i)}$ represent design matrix coding of the row (column) variable, such as effect or dummy coding. In the case of dummy coding, the $D_{R(i)} = 1$ when $R = i$ and zero otherwise. The degrees of freedom are $(I-1)(J-2)$ for this model. The Row Effects model is invariant to the orderings of the rows, but the choice of scores for the columns will effect the results. An additional advantage of this model is that the $\phi_{R(i)}$ parameters may be used as a guide to scoring the row variable when the column scores are chosen properly. For example, if the scores follow an ordering that is somewhat evenly spaced, integer scoring may be appropriate. Additionally, two categories with approximately the same scores might be combined into one category. The same model can be constructed for column effects using $\phi_{C(j)}D_{C(j)}x_j$ instead of the $\phi_{R(i)}D_{R(i)}$.

This approach may also be extended to the case when both variables are nominal, or to verify a chosen scoring scheme for the two variables. The Row and Column (R+C) effects model is

$$\log F_{ij} = \lambda + \lambda_i^R + \lambda_j^C + \sum_{i=1}^{I-1} \phi_{R(i)} D_{R(i)} y_j + \sum_{j=1}^{J-1} \phi_{C(j)} D_{C(j)} x_i.$$

The degrees of freedom for this model are $(I-2)(J-2)$. The scores estimated are dependent on the scoring chosen for the model because they appear in both sets of association terms. Thus the fit of this model is not invariant to the category orderings, unlike the row and column effects models. If your main purpose is to score both the row and column variables, then a log-multiplicative model (RC model) that estimates these scores should be fit to the data. Currently, SAS (or any of the major statistical packages that I'm aware of) does not have a procedure to fit these models directly. However, when the column scores are appropriately chosen, the estimated row scores will be very similar to those of the RC model (they will be the same, up to a linear transformation, if the column scores from the RC model are used for the row effects model).

3 SAS DATA STEP

The data need to be explicitly coded in both quantitative and qualitative terms to correctly compute all of the possible association terms. Each variable will be given a score (x_i or y_j) and a corresponding set of dummy (or effect) coded variables. The dummy variables are used to compute the λ parameters for the main effects (marginals), and the scores are used to get the association terms used when estimating the ϕ parameters. An example

with an 11 X 6 table of Occupation by Years-of-Schooling is used to illustrate the models in SAS. These data are from the 1988 General Social Survey and are discussed in Clogg & Shihadeh (1994). The data table is shown at the bottom of the SAS DATA step. The rows represent the eleven occupations, listed in Table 2 below. The columns are the six categories of Years-of-Schooling: 0-7, 8-11, 12 (high school graduate) 13-15, 16 (college graduate) and 17+.

```
DATA GSS88;
DO OCCUP = 1 TO 11; * INTEGER SCORING;
DO EDUC = 1 TO 6;
INPUT COUNT @@;
LLASSOC=OCCUP*EDUC;
ARRAY OCCARR[*] OCCUP1-OCCUP10;
ARRAY RASSOC[*] RASSOC1-RASSOC10;
DO I= 1 TO 10;
IF OCCUP = I THEN OCCARR[I] = 1;
ELSE OCCARR[I] = 0;
RASSOC[I] = OCCARR[I]*EDUC;
END;
ARRAY EDARR[*] EDUC1-EDUC5;
ARRAY CASSOC[*] CASSOC1-CASSOC5;
DO J= 1 TO 5;
IF EDUC = J THEN EDARR[J] = 1;
ELSE EDARR[J] = 0;
CASSOC[J] = EDARR[J]*OCCUP;
END;
OUTPUT;
END;
DROP I J;
CARDS;
0 2 17 27 34 48
0 1 13 16 36 27
5 21 55 67 36 19
2 23 101 80 15 5
3 36 43 27 3 3
0 3 13 11 3 1
10 34 31 13 2 1
5 14 27 15 4 1
1 3 4 4 1 0
7 36 61 39 9 4
2 16 8 13 6 3
;
```

The WORK.GSS88 data set has one line for each cell in the table (66), and 34 variables. The OCCUP and EDUC variables are the integer scores, while the LLASSOC (Linear X Linear Association) variable is their product. The OCCUP1, OCCUP2, ..., OCCUP10 variables are the ten dummy codes for the eleven occupations, and the EDUC1,

EDUC, ..., EDUC5 represent the dummy codes for the six education levels. RASSOC1, RASSOC2, ..., RASSOC10 are the ten Row Association variables computed as the product of the dummy codes for Occupation (row) and the scores for Years-of-Schooling (column). The column association variables CASSOC1, CASSOC2, ..., CASSOC5 are computed similarly. Thus the RASSOC i variables correspond to the $D_{R(i)}y_j$, and the CASSOC j correspond to the $D_{C(j)}x_i$.

4 PROC GENMOD

The PROC GENMOD code is presented in the order that the models are discussed above. Class statements are used to save some typing on the model statement because they cause the program to automatically dummy code the variables.

```
TITLE 'INDEPENDENCE MODEL';
PROC GENMOD;
CLASS OCCUP EDUC;
MODEL COUNT = OCCUP EDUC
  /DIST=POI LINK=LOG TYPE3;
```

```
TITLE 'UNIFORM ASSOCIATION MODEL';
PROC GENMOD;
CLASS OCCUP EDUC;
MODEL COUNT = OCCUP EDUC LASSOC
  /DIST=POI LINK=LOG TYPE3;
RUN;
```

```
TITLE 'ROW ASSOCIATION MODEL';
PROC GENMOD;
CLASS OCCUP EDUC;
MODEL COUNT =
  OCCUP EDUC RASSOC1-RASSOC10
  /DIST=POI LINK=LOG TYPE3;
RUN;
```

```
TITLE 'COLUMN ASSOCIATION MODEL';
PROC GENMOD;
CLASS OCCUP EDUC;
MODEL COUNT =
  OCCUP EDUC CASSOC1-CASSOC5
  /DIST=POI LINK=LOG TYPE3;
RUN;
```

```
TITLE 'ROW+COLUMN ASSOCIATION MODEL';
PROC GENMOD;
CLASS EDUC OCCUP;
MODEL COUNT =
  OCCUP EDUC RASSOC1-RASSOC10
  CASSOC1-CASSOC5
```

```
/DIST=POI LINK=LOG TYPE3;
RUN;
```

5 RESULTS

The results show that there is a substantial reduction in the likelihood-ratio statistic when the Uniform Association model is compared to the Independence model. Similarly, the Row Effects model is a significant improvement over the Uniform Association model. The improvements for the Column Effects and Row+Column Effects models are not very large.

Table 1

Model	df	LR
Independence	50	455.06
Uniform Association	49	240.60
Row Effects	40	62.49
Column Effects	45	215.14
Row+Column	36	54.24

If we accept the Row Effects model, the one-degree-of-freedom parameter estimates (i.e., the $\phi_{R(i)}$) show how the occupations may be ordered to correspond to education. Table 2 shows that the two professional occupations may be combined into one score, and similarly for the Operatives, Farmers and Service occupations. The respondents in the missing category should be placed in the middle of the occupations. The vector of parameter estimates is sometimes normalized to have a mean of zero and a length of one when presented.

Table 2

Occupation	Variable	Score	Std Err
Professional-I	RASSOC1	1.27	0.17
Professional-II	RASSOC2	1.23	0.18
Managers	RASSOC3	0.44	0.15
Clerical	RASSOC4	0.13	0.15
Craftsmen-I	RASSOC5	-0.27	0.17
Craftsmen-II	RASSOC6	0.22	0.21
Operatives (NT)	RASSOC7	-0.66	0.18
Operatives	RASSOC8	-0.24	0.18
Farmers	RASSOC9	-0.20	0.30
Service	RASSOC10	-0.16	0.16
Missing	(Fixed)	0.00	

6 CONCLUSION

The association models shown here are "in-between" the independence and saturated log-linear models. However, they may capture associations in the data that would not be found in an ordinary log-linear model. The flexibility of the SAS DATA step and the ease of PROC GENMOD make fitting these models very straightforward. These features also allow for adding a third "layer" or grouping variable. For instance, the gender of the respondent may be added to the table above, or similar tables may be constructed over time. When grouping variables are used, the association parameters in the two-way tables may be constrained to be the same or allowed to vary across groups. The ordinal structure of layer variables can be exploited by modeling a linear or other trend of association over time.

7 REFERENCES

Clogg, C. C., and Shihadeh, E. S. (1994). *Statistical Models for Ordinal Variables*. Thousand Oaks: Sage.

Goodman, L. A. (1979). Simple models for the association in cross classifications having ordered categories. *Journal of the American Statistical Association*, 74, 537-552.

SAS is a registered trademark of SAS Institute Inc. in the USA and other countries.

® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

James R. Hershey
Department of Psychology
University of Illinois at Urbana-Champaign
603 East Daniel Street
Champaign IL 61820
(217) 333-6750
jhershey@uiuc.edu