

USING SAS GENMOD PROCEDURE TO FIT DIAGONAL CLASS MODELS TO SQUARE CONTINGENCY TABLES HAVING ORDERED CATEGORIES

by

H. Bayo Lawal
Department of Statistics
St. Cloud State University
St. Cloud, MN 56301, USA.

ABSTRACT

We shall demonstrate in this paper, how the generalized linear model procedure GENMOD in SAS can be utilized to implement the class of models that can be fitted by Lawal and Upton's (1995) FORTRAN algorithm to square tables having ordered classificatory variables. Composite models as discussed by Lawal and Upton will also be fitted to the tables. Applications to 4×4 and 5×5 square tables will be discussed. Parameter estimates as well as the relevant odds ratios will also be discussed from the SAS output.

1. INTRODUCTION

Let n_{ij} denote the observed frequency in the i -th row and j -th column of an $I \times I$ square contingency table ($i, j = 1, 2, \dots, I$), and let m_{ij} denote the corresponding expected frequency under some model. If we assume here that a multinomial sampling scheme applies to the $I \times I$ table, though other sampling schemes may also be assumed without loss of generality. If θ_{ij} denote the $(I-1)(I-1)$ local odds-ratio formed from adjacent 2×2 subtables, then

$$\theta_{ij} = (m_{ij}m_{i+1,j+1})/(m_{i,j+1}m_{i+1,j}).$$

2. THE PRINCIPAL DIAGONAL CLASS

The principal diagonal class models has:

$$\theta_{ij} = \begin{cases} \theta_i & \text{if } (i = j) \\ 1 & \text{if } (i \neq j) \end{cases}$$

The fixed distance model (F) with constant or fixed distance parameter δ for an $I \times I$ table can be described in terms of its expected values as:

$$m_{ij} = \alpha_i \beta_j \delta^k \quad (1)$$

where $k = |i - j|$ for $(i, j) = 1, 2, \dots, I$. Under this model, the estimated odds-ratios are estimated as:

$$\hat{\theta}_{ij} = \begin{cases} \exp\{-\hat{\delta}\} & \text{if } i = j \\ 1 & \text{if } i \neq j \end{cases}$$

The model can be implemented in SAS procedure GENMOD to the 5×5 British Social Mobility (Glass, 1954) table below by defining a quantitative variable F in the CLASS command line of GENMOD. The variable F is defined as:

$$F = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 1 & 2 & 3 & 4 \\ 1 & 1 & 1 & 2 & 3 \\ 1 & 1 & 1 & 1 & 2 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Occupational status for British father-son pairs

Father's Status	Sons' Status				
	(1)	(2)	(3)	(4)	(5)
(1)	50	45	8	18	8
(2)	28	174	84	154	55
(3)	11	78	110	223	96
(4)	14	150	185	714	447
(5)	3	42	72	320	411

The model is based on $I^2 - [1 + 2(I - 1) + 1] = I(I - 2)$ degrees of freedom.

THE VARIABLE DISTANCE MODEL:

The variable distance model (V) for an $I \times I$ table is also a member of this class and can be written in log-linear terms as follows:

$$\log(m_{ij}) = \mu + \lambda_i^R + \lambda_j^C + \lambda_{ij}^D \quad (2)$$

for $(i, j) = 1, 2, \dots, I$ and where for some δ_k , $1 \leq k \leq (I - 1)$,

$$\lambda_{ij}^D = \begin{cases} \sum_{k=1}^{j-1} \delta_k & \text{if } i < j \\ 0 & \text{if } i = j \\ \sum_{k=j}^{i-1} \delta_k & \text{if } j < i \end{cases}$$

There are thus $I - 1$ such distance parameters, namely $\delta_1, \delta_2, \dots, \delta_{I-1}$, which represent respectively, distances from category 1 to 2, 2 to 3, ..., and $(I - 1)$ to I . The model is based on $I^2 - [2(I - 1) + I] = (I - 1)(I - 2)$ degrees of freedom and the corresponding local odds-ratios for the model are also given by:

$$\theta_{ij} = \begin{cases} \exp\{-2\delta_i\} & \text{if } i = j \\ 1 & \text{if } i \neq j \end{cases}$$

For the 5×5 table, there are 4 of such parameters and they can be implemented in SAS by defining the following four factor variables:

$$D1 = \begin{bmatrix} 2 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 2 \\ 1 & 2 & 2 & 2 & 2 \\ 1 & 2 & 2 & 2 & 2 \\ 1 & 2 & 2 & 2 & 2 \end{bmatrix}, \quad D2 = \begin{bmatrix} 3 & 3 & 1 & 1 & 1 \\ 3 & 3 & 1 & 1 & 1 \\ 1 & 1 & 3 & 3 & 3 \\ 1 & 1 & 3 & 3 & 3 \\ 1 & 1 & 3 & 3 & 3 \end{bmatrix},$$

$$D3 = \begin{bmatrix} 4 & 4 & 4 & 1 & 1 \\ 4 & 4 & 4 & 1 & 1 \\ 4 & 4 & 4 & 1 & 1 \\ 1 & 1 & 1 & 4 & 4 \\ 1 & 1 & 1 & 4 & 4 \end{bmatrix}, \quad D4 = \begin{bmatrix} 5 & 5 & 5 & 5 & 1 \\ 5 & 5 & 5 & 5 & 1 \\ 5 & 5 & 5 & 5 & 1 \\ 5 & 5 & 5 & 5 & 1 \\ 1 & 1 & 1 & 1 & 5 \end{bmatrix}.$$

The parameter estimated under both models F and V are respectively: $F: \hat{\delta} = -1.0672$ and $\hat{\delta}_1 = -1.2715; \hat{\delta}_2 = -0.6419; \hat{\delta}_3 = -0.4158; \hat{\delta}_4 = -0.4594$; and the corresponding estimated odds-ratios are displayed below:

$$F: \hat{\theta}_{ij} = \begin{bmatrix} 2.907 & 1.000 & 1.000 & 1.000 \\ 1.000 & 2.907 & 1.000 & 1.000 \\ 1.000 & 1.000 & 2.907 & 1.000 \\ 1.000 & 1.000 & 1.000 & 2.907 \end{bmatrix}, \quad V: \hat{\theta}_{ij} = \begin{bmatrix} 12.719 & 1.000 & 1.000 & 1.000 \\ 1.000 & 3.610 & 1.000 & 1.000 \\ 1.000 & 1.000 & 2.297 & 1.000 \\ 1.000 & 1.000 & 1.000 & 2.506 \end{bmatrix}$$

3. THE DIAGONAL BAND MODELS:

This class of models has

$$\theta_{ij} = 1 \quad \text{unless} \quad |i - j| \leq 1$$

THE UNIFORM LOYALTY MODEL:

The uniform loyalty model (L) is a member of this class of models and its expected values can be expressed as:

$$m_{ij} = \begin{cases} \alpha_i \beta_j & (i \neq j) \\ \lambda \alpha_i \beta_j & (i = j) \end{cases} \quad (3)$$

We can implement this model in SAS by defining a quantitative variable for a 5×5 table as:

$$L = \begin{bmatrix} 2 & 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 & 2 \end{bmatrix}$$

The model is based on $I^2 - [1 + 2(I - 1) + 1] = I(I - 2)$ d.f.

THE TRIANGLE (T) MODEL:

Goodman(1986) described the non-independence triangle (T) model

$$m_{ij} = \alpha_i \beta_j \gamma_{ij} \quad (4)$$

where

$$\gamma_{ij} = \begin{cases} \tau_1 & \text{for } i > j \\ 1 & \text{for } i = j \\ \tau_2 & \text{for } i < j \end{cases}$$

The τ parameters pertain to the upper-right triangle and lower left triangles in the square table.

This model is implemented in SAS by defining a factor variable T as follows:

$$T = \begin{bmatrix} 1 & 3 & 3 & 3 & 3 \\ 2 & 1 & 3 & 3 & 3 \\ 2 & 2 & 1 & 3 & 3 \\ 2 & 2 & 2 & 1 & 3 \\ 2 & 2 & 2 & 2 & 1 \end{bmatrix}$$

The model is based on $I^2 - [1 + 2(I - 1) + 2] = (I^2 - 2I - 1)$ degrees of freedom.

THE QUASI-INDEPENDENCE Q MODEL:

The model of quasi-independence (Q) is described in terms of its expected values as:

$$m_{ij} = \alpha_i \beta_j \psi_{ij} \quad (5)$$

where $\psi_{ij} = 1$ for $i \neq j$.

The SAS implementation of this model calls for defining a factor variable Q as:

$$Q = \begin{bmatrix} 2 & 1 & 1 & 1 & 1 \\ 1 & 3 & 1 & 1 & 1 \\ 1 & 1 & 4 & 1 & 1 \\ 1 & 1 & 1 & 5 & 1 \\ 1 & 1 & 1 & 1 & 6 \end{bmatrix}$$

The model is based on $I^2 - [1 + 2(I - 1) + I] = (I^2 - 3I + 1)$ degrees of freedom and is sometimes referred to as the mover-stayer model. Corresponding estimated local odds-ratios are similarly displayed below. The general structures of these odds-ratios are given in Lawal (1996).

$$L : \hat{\theta}_{ij} = \begin{bmatrix} 3.838 & 0.510 & 1.000 & 1.000 \\ 0.510 & 3.838 & 0.510 & 1.000 \\ 1.000 & 0.510 & 3.838 & 0.510 \\ 1.000 & 1.000 & 0.510 & 3.838 \end{bmatrix}, \quad T : \hat{\theta}_{ij} = \begin{bmatrix} 3.875 & 0.615 & 1.000 & 1.000 \\ 0.420 & 3.875 & 0.615 & 1.000 \\ 1.000 & 0.420 & 3.875 & 0.615 \\ 1.000 & 1.000 & 0.420 & 3.875 \end{bmatrix}$$

$$Q : \hat{\theta}_{ij} = \begin{bmatrix} 131.367 & 0.248 & 1.000 & 1.000 \\ 0.248 & 6.846 & 0.589 & 1.000 \\ 1.000 & 0.589 & 1.671 & 1.017 \\ 1.000 & 1.000 & 1.017 & 2.798 \end{bmatrix}$$

4. THE FULL DIAGONAL MODELS:

Goodman (1972, 1979, 1986) introduced the following models:

THE DIAGONAL (D) MODEL:

The diagonal (D) model has its expected values described by:

$$m_{ij} = \alpha_i \beta_j \delta_k \text{ for } (i \neq j).$$

where $k = i - j$ and $\delta_0 = 1$

The implementation of this model in SAS can be accomplished by defining a factor variable D which can take any of the following forms- although, the latter form is often preferred.

$$D = \begin{bmatrix} 1 & -1 & -2 & -3 & -4 \\ 1 & 1 & -1 & -2 & -3 \\ 2 & 1 & 1 & -1 & -2 \\ 3 & 2 & 1 & 1 & -1 \\ 4 & 3 & 2 & 1 & 1 \end{bmatrix}, \text{ or } D = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 6 & 1 & 2 & 3 & 4 \\ 7 & 6 & 1 & 2 & 3 \\ 8 & 7 & 6 & 1 & 2 \\ 9 & 8 & 7 & 6 & 1 \end{bmatrix}$$

The model is based on $I^2 - [1 + 2(I - 1) + (2I - 3)] = (I - 2)^2$ degrees of freedom.

THE DIAGONALS-ABSOLUTE (DA) MODEL:

$$m_{ij} = \alpha_i \beta_j \delta_k \text{ for } (i \neq j).$$

where $k = |i - j|$ and $i \neq j$.

$$DA = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 1 & 2 & 3 & 4 \\ 3 & 2 & 1 & 2 & 3 \\ 4 & 3 & 2 & 1 & 2 \\ 5 & 4 & 3 & 2 & 1 \end{bmatrix}$$

The model is based on $I^2 - [1 + 2(I - 1) + (I - 1)] = (I - 1)(I - 2)$ degrees of freedom.

THE UNIFORM ASSOCIATION (U) MODEL:

The uniform association model U belongs to this class of models and has $\theta_{ij} = \theta$ for $(i, j) = 1, 2, \dots, (I - 1)$ and with expected values expressed as:

$$m_{ij} = \alpha_i \beta_j \theta^{ij}.$$

For this model all local odds-ratios are equal and there is said to be a uniform local association for all cells in the square table. The model has $I(I-2)$ d.f and is sometimes referred to as the *uniform diagonals* model.

This model can also be implemented in SAS by defining a quantitative variable $U = i * j$ for $(i, j) = 1, 2, \dots, 5$ which reduces to:

$$U = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 4 & 6 & 8 & 10 \\ 3 & 6 & 9 & 12 & 15 \\ 4 & 8 & 12 & 16 & 20 \\ 5 & 10 & 15 & 20 & 25 \end{bmatrix}$$

The model is based on $I^2 - [1 + 2(I - 1) + 1] = I(I - 2)$ degrees of freedom. We also give below the estimated odds-ratios for the full diagonal class models.

$$D : \hat{\theta}_{ij} = \begin{bmatrix} 2.197 & 1.323 & 1.317 & 1.040 \\ 0.998 & 2.197 & 1.323 & 1.317 \\ 1.650 & 0.998 & 2.197 & 1.323 \\ 1.672 & 1.650 & 0.998 & 2.197 \end{bmatrix}, \quad DA : \hat{\theta}_{ij} = \begin{bmatrix} 2.182 & 1.162 & 1.464 & 1.218 \\ 1.162 & 2.182 & 1.162 & 1.464 \\ 1.464 & 1.162 & 2.182 & 1.162 \\ 1.218 & 1.464 & 1.162 & 2.182 \end{bmatrix}$$

$$U : \hat{\theta}_{ij} = \begin{bmatrix} 1.510 & 1.510 & 1.510 & 1.510 \\ 1.510 & 1.510 & 1.510 & 1.510 \\ 1.510 & 1.510 & 1.510 & 1.510 \\ 1.510 & 1.510 & 1.510 & 1.510 \end{bmatrix}$$

The following commands in SAS which utilizes procedure GENMOD will implement all the models described in the previous sections.

```

L:  PROC GENMOD; CLASS R C;
      MODEL COUNT= R C L/DIST=POI LINK=LOG;
F:  PROC GENMOD; CLASS R C;
      MODEL COUNT= R C F/DIST=POI LINK=LOG;
U:  PROC GENMOD; CLASS R C;
      MODEL COUNT= R C U/DIST=POI LINK=LOG;
V:  PROC GENMOD; CLASS R C D1-D4;
      MODEL COUNT= R C D1-D4/DIST=POI LINK=LOG;
Q:  PROC GENMOD; CLASS R C Q;
      MODEL COUNT= R C Q/DIST=POI LINK=LOG;
D:  PROC GENMOD; CLASS R C D;
      MODEL COUNT= R C D/DIST=POI LINK=LOG;
DA: PROC GENMOD; CLASS R C DA;
      MODEL COUNT= R C DA/DIST=POI LINK=LOG;
T:  PROC GENMOD; CLASS R C T;
      MAKE 'OBSTATS' OUT=TRI;
      MODEL COUNT= R C T/DIST=POI LINK=LOG;
      PROC PRINT; DATA=TRI;
      VAR PRED;

```

The print procedure above for example will print both the observed and expected cell counts under model T.

5. COMPOSITE MODELS:

Each of the models described above may not alone be able to completely explain the complex variations in any given square table we are likely to encounter. Consequently, it is often appropriate to consider second order combinations of each of the models in the preceding sections. It has been found that such a combination(s) usually explains better the data (Lawal & Upton, 1989); (Lawal, 1992); Goodman (1972, 1986). For instance, the diagonals- absolute triangle (DAT) model (Goodman, 1986) is a combination of the DA and T models. This model can be implemented in SAS with the following statements:

```
PROC GENMOD; CLASS R C DA T;
MODEL COUNT= R C DA T/DIST=POI LINK=LOG;
```

Other composite models can also be formed by noting that any of the variables D1-D4, Q, DA, T must always be specified in the CLASS line of procedure GENMOD and that U, L and F need not be specified in the CLASS line, since these are considered as quantitative variables.

However, there is an hierarchy of complexity to the constraints imposed by the individual models Lawal (1993). Thus:

$$D, U \subseteq D$$

$$D, T \subseteq D$$

$$DA \subseteq D$$

For higher order model combinations, we note for example that model VUQDDAT will be equivalent to model QVD and that for instance, model QD would be based when $I \geq 4$ on $(I^2 - 5I + 5)$ d.f.

Similarly, model QDT is equivalent, since $T \subseteq D$, to model QD.

For 4×4 Tables, the corresponding factor and quantitative variables necessary to implement these models in SAS are defined below as:

$$DA = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 1 & 2 & 3 \\ 6 & 5 & 1 & 2 \\ 7 & 6 & 5 & 1 \end{bmatrix}, \quad U = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 6 & 8 \\ 3 & 6 & 9 & 12 \\ 4 & 8 & 12 & 16 \end{bmatrix}$$

$$Q = \begin{bmatrix} 2 & 1 & 1 & 1 \\ 1 & 3 & 1 & 1 \\ 1 & 1 & 4 & 1 \\ 1 & 1 & 1 & 5 \end{bmatrix}, \quad T = \begin{bmatrix} 1 & 3 & 3 & 3 \\ 2 & 1 & 3 & 3 \\ 2 & 2 & 1 & 3 \\ 2 & 2 & 2 & 1 \end{bmatrix}, \quad L = \begin{bmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{bmatrix}$$

$$F = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 1 & 2 & 3 \\ 1 & 1 & 1 & 2 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \quad D1 = \begin{bmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 1 & 2 & 2 & 2 \\ 1 & 2 & 2 & 2 \end{bmatrix}$$

$$D2 = \begin{bmatrix} 3 & 3 & 1 & 1 \\ 3 & 3 & 1 & 1 \\ 1 & 1 & 3 & 3 \\ 1 & 1 & 3 & 3 \end{bmatrix}, \quad D3 = \begin{bmatrix} 4 & 4 & 4 & 1 \\ 4 & 4 & 4 & 1 \\ 4 & 4 & 4 & 1 \\ 1 & 1 & 1 & 4 \end{bmatrix}$$

REFERENCES

- Glass, D. V. (1954) *Social Mobility in Britain*, Glenco Ill: Free Press.
- Goodman, L. A. (1972). Some multiplicative models for the analysis of cross-classified data. In Le Carm *et al.*, *Proceesdings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, 649-695. Berkeley: University of California Press.
- Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *J. Amer. Statist. Assoc.*, 74,537-552.
- Goodman, L. A. (1986). Some Useful Extensions of the usual Correspondence Analysis Approach and the Usual Log-linear Models approach in the Analysis of Contingency Tables with discussion. *International Statistical Review*, 54, 243-270.
- Haberman, S. (1974). *The Analysis of Frequency Data*. Chicago: The University Press:
- Haberman, S. (1978). *Analysis of Quantitative Data, Vol. 2*, Academic Press. Inc., New York.
- Lawal, H. B. (1992). Parsimonious uniform-distance association models for the occupational mobility data. *J. Japan Statist. Soc.*, 22, 183-192.

- Lawal, H. B. (1993). Association, Symmetry and Diagonal models for occupational mobility and other similar square contingency tables having ordered categorical variables. *Biometrical Journal*, 35, 193-206.
- Lawal, H. B. & Upton, G.J.G. (1990). Alternative interaction structures in square contingency tables having ordered classificatory variables. *J. Quality and Quantity*, 24, 107-127.
- Lawal, H. B. & Upton, G.J.G. (1995). An algorithm for fitting models to $N \times N$ square contingency tables having ordered categories. *Comm. Statist.* 24, (3) forthcoming.