MODEL SELECTION FOR FOUR- AND HIGHER-DIMENSIONAL TABLES

M. A. Toleman

As the number of dimensions of the contingency table increases so does the number of possible hierarchical log-linear models. For example, for three dimensions there are eight models and for four dimensions there are 113 models (Bishop et al, 1975). Obviously we cannot test the fit of all these models so we need a method of selecting terms to be included in our "best fit" model. There is no best method for selecting a log-linear model just as there is no best method for variable selection in multiple regression. Three methods of model selection are discussed, each in the context of a single data set. The program GLIM is used throughout to fit models.

1. DATA SET FOR ILLUSTRATION

Table 1 represents hypothetical data for 564 steers categorized by four variables, (1) breed, (2) horn length pre-dehorning, (3) dehorning instrument and (4) regrowth at six months post-dehorning. There are two breeds, brahman and sahiwal, three categories of horn length, less than 2.5 cm, 2.5-3.5 cm and greater than 3.5 cm, two dehorning instruments, scoop dehorner and large hodge pattern calf dehorner, and two categories of regrowth, nil and some regrowth.

Table 1 Frequency data on breed x horn length pre-dehorning x dehorning instrument x regrowth at six months post-dehorning.

Breed	Horn length pre-dehorning	Dehorning Instrument: Scoop		Hodge	
		Regrowth			
		Nil	Some	Nil	Some
Brahman	≤ 2.5 cm	54	44	25	18
	2.5-3.5 cm	77	63	64	21
	≥ 3.5 cm	22	25	13	5
Sahiwal	≤ 2.5 cm	13	5	18	1
	2.5-3.5 cm	21	4	33	5
	≥ 3.5 cm	14	3	11	5

2. BROWN'S MODEL SELECTION METHOD

In the analysis of continuous non-orthogonal data the significance of adding an effect depends on what effects are currently in the model. The same holds for the analysis of contingency tables. We cannot look at the effect of adding a particular term to all models. Hence, a model selection method similar to that in Brown (1976) relies on the calculation of two test statistics for each effect. For example, to test for the term u_{12} in the example of Section 1 four models would be considered

[1][2][3][4] (2.1)

[12][3][4] (2.2)

[12][13][14][23][24][34] (2.3)

[13][14][23][24][34] (2.4)

Each model has an associated residual deviance (G^2) which follows approximately a χ^2 distribution with degrees of freedom given by

d.f. = number of cells - number of parameters fitted.

The change in deviance between models (2.2) and (2.1) is the effect of including u_{12} in the "main effects only" model ($+\Delta G^2$) while the change in deviance between models (2.4) and (2.3) is the effect of excluding u_{12} from the "all two-factor effects" model ($-\Delta G^2$). Both test statistics follow approximately χ^2 distributions and have the same degrees of freedom as given for effects in analysis of variance. Each effect should give some idea of the relative importance of the effect u_{12}. In a similar way the other two-factor effects and the three factor effects may be examined.

Table 2 gives the ΔG^2 values for this method for the example.

Table 2 ΔG^2 values for data in Table 1.

u term	d.f.	$+\Delta G^2$	$-\Delta G^2$
u_{12}	2	6.36*	8.60*
u_{13}	1	18.52**	13.66**
u_{14}	1	26.85**	22.09**
u_{23}	2	3.73	3.90
u_{24}	2	2.31	2.57
u_{34}	1	16.62**	9.79**
u_{123}	2	0.01	0.00
u_{124}	2	0.39	0.45
u_{134}	1	0.49	0.45
u_{234}	2	1.19	1.12

* $P < 0.05$

**$P < 0.01$

Effects that warrant inclusion in a model are u_{12}, u_{13}, u_{14} and u_{34} since they are significant ($P \leq 0.05$) for both test statistics. For some data sets one test statistic from an effect may be significant and the other not. In this case models should be considered with and without this effect. Neither adding extra terms nor excluding any of these terms enhances the fit so we have the model

[12][13][14][34] (2.5)

with $G^2 = 15.88$ on 13 d.f. ($P > 0.05$) as the best model by this method.

3. STEPWISE SELECTION METHODS

The methods outlined here are similar to those used in the analysis of continuous data with unequal sub-class numbers. Goodman (1971) first proposed their use for log-linear models. Fienberg (1977, p. 65-66) describes the process in detail.

There are two approaches (a) forward selection or (b) backward elimination. Both methods will be illustrated using the data of Table 1. The significance level used throughout is 0.05.

(a) Forward Selection: Fit the models

	d.f.	G^2	
[1] [2] [3] [4]	18	78.38*	(3.1)
[12] [13] [14] [23] [24] [34]	9	8.44	(3.2)
[123] [124] [134] [234]	2	6.38*	(3.3)

Since (3.1) does not fit and (3.2) does fit we look for a model between these. (Note that (3.3) does not fit well. Goodman ignores this situation).

Add each two-factor term to (3.1) to get the following models

	d.f.	G^2	ΔG^2(from (3.1))	d.f.	
[3] [4] [12]	16	72.02	6.36	2	(3.4)
[2] [4] [13]	17	59.86*	18.52*	1	(3.5)
[2] [3] [14]	17	51.53*	26.85*	1	(3.6)
[1] [4] [23]	16	74.65*	3.73	2	(3.7)
[1] [3] [24]	16	76.07*	2.31	2	(3.8)
[1] [2] [34]	17	61.76*	16.62*	1	(3.9)

Since u_{14} is the most significant effect add it to (3.1). Now adding each two-factor term to (3.6) we get the following models

	d.f.	G^2	ΔG^2(from (3.6))	d.f.	
[3] [4] [12]	15	45.16*	6.37*	2	(3.10)
[2] [14] [13]	16	33.00*	18.53*	1	(3.11)
[14] [23]	15	47.79*	3.74	2	(3.12)
[3] [14] [24]	15	49.21*	2.32	2	(3.13)
[2] [14] [34]	16	34.91*	16.62*	1	(3.14)

Since u_{13} is the most significant effect add it to (3.6). Neither u_{14} nor u_{13} may be deleted at this stage. Now adding each two-factor term to (3.11) we get the following models

	d.f.	G^2	ΔG^2(from (3.11)	d.f.	
[14] [13] [12]	14	26.64*	6.36*	2	(3.15)
[14] [13] [23]	14	29.27*	3.73	2	(3.16)
[14] [13] [24]	14	30.69*	2.31	2	(3.17)
[14] [13] [34]	15	22.15	10.85*	1	(3.18)

Since u_{34} is the most significant effect add it to (3.11). None of u_{14}, u_{13} or u_{34} may be deleted at this stage. Now adding each two-factor term to (3.18) we get the following models

	d.f.	G^2	ΔG^2(from (3.18))	d.f.	
[14] [13] [34] [12]	13	15.88	6.27*	2	(3.19)
[14] [13] [34] [23]	13	18.51	3.64	2	(3.20)
[14] [13] [34] [24]	13	19.93	2.22	2	(3.21)

Since u_{12} is the only significant effect we add it to (3.18). None of u_{14}, u_{13}, u_{34} or u_{12} may be deleted at this stage. Now adding each two-factor term to (3.19) we get the following models

	d.f.	G^2	ΔG^2(from (3.19))	d.f.
[14] [13] [34] [12] [23]	11	12.34	3.54	2
[14] [13] [34] [12] [24]	11	11.01	4.87	2

Neither model suggests an improved fit so our best model is given by (3.19)

(b) Backward elimination: Fit the models (3.1), (3.2) and (3.3). This time we start with (3.2) and subtract two-factor terms to get the following models

	d.f.	G^2	ΔG^2(from (3.2))	d.f.	
[13] [14] [23] [24] [34]	11	17.03	8.59*	2	(3.22)
[12] [14] [23] [24] [34]	10	22.10*	13.66*	1	(3.23)
[12] [13] [23] [24] [34]	10	30.53*	22.09*	1	(3.24)
[12] [13] [14] [24] [34]	11	12.34	3.90	2	(3.25)
[12] [13] [14] [23] [34]	11	11.01	2.57	2	(3.26)
[12] [13] [14] [23] [24]	10	18.23	9.79*	1	(3.27)

Since deleting u_{24} is the most non-significant event choose (3.26) and subtract two-factor effects to get the following models

	d.f.	G^2	ΔG^2(from (3.26))	d.f.	
[13] [14] [23] [34]	13	18.51	7.50*	2	(3.28)
[12] [14] [23] [34]	12	24.65*	13.64*	1	(3.29)
[12] [13] [23] [34]	12	32.01*	21.00*	1	(3.30)
[12] [13] [14] [34]	13	15.88	4.87	2	(3.31)
[12] [13] [14] [23]	12	21.77*	10.76*	1	(3.32)

Since deleting u_{23} is the most non-significant event choose (3.31). Neither u_{24} nor u_{23} can be added back in at this stage. Now subtracting each two-factor effect from (3.31) we get the following models

	d.f.	G^2	ΔG^2(from (3.31))	d.f.
[13] [14] [34]	15	22.15	6.27*	2
[12] [14] [34]	14	28.54*	12.66*	1
[12] [13] [34]	14	36.87*	20.99*	1
[12] [13] [14]	14	26.64*	10.76*	1

Deleting any term would now significantly affect the fit so our model is (3.31). No deleted terms can be added back in to improve the fit so our "best" model is (3.31).

Both (a) and (b) give the same "best" model.

4. METHOD OF STANDARDIZED PARAMETER ESTIMATES

This method relies on some simple formulae for the asympotic variance of the estimated u-terms in the full log-linear model (Fienberg, 1977). Each u-term is expressible as a linear combination of the logarithms of the expected cell values. For example in a four dimensional table we have terms like

$$u_{1234(ijkl)} = \sum_{i,j,k,l} \beta_{ijkl} \log m_{ijkl}$$

where $\sum_{i,j,k,l} \beta_{ijkl} = 0$

The maximum likelihood estimate for this term is given by

$$\hat{u}_{1234(ijkl)} = \sum_{i,j,k,l} \beta_{ijkl} \log x_{ijkl}$$

with variance given by

$$\sum_{i,j,k,l} \beta^2_{ijkl} \quad x^{-1}_{ijkl}$$

Bishop et al (1975) give the following general formulae for the above:

$$u_{\theta(j)} = \sum_i \beta_i \log m_i$$

where $\sum_i \beta_i = 0$, with maximum likelihood estimate

$$\hat{u}_{\theta(j)} = \sum_i \beta_i \log x_i$$

and variance given by

$$V(\hat{u}_{\theta(j)}) = \sum_i \beta_i^2 x_i^{-1}$$

Standardized parameter estimates are then obtained as the estimated u-term divided by the estimated standard deviation for that term.

Note that cells with zero values invalidate this method since log 0 is undefined.

Note also that the parameter estimates given by the constraints used in GLIM are not appropriate for this method. However, it is possible to construct our own constraints and still use GLIM.

To select a "base" model these standardized parameter estimates are tested against a z-score, at an appropriate significance level. Further models are then analysed using this "base" model and the stepwise procedures of Section 3.

Table 4 gives the parameter estimates, estimated standard errors and standardized parameter estimates for the data from Table 1.

Table 4 Parameter estimates, estimated standard errors and standardized parameter estimates for data from Table 1.

Effect	Estimate	S.E.	z-score
u_1	0.6546	0.0555	11.79**
u_L	-0.1699	0.0680	- 2.50*
u_Q	-0.2350	0.0392	- 5.99**
u_3	0.2259	0.0555	4.07**
u_4	0.5165	0.0555	9.31**
u_{1L}	-0.2549	0.0680	- 3.75**
u_{1Q}	-0.0573	0.0392	- 1.46
u_{13}	0.1977	0.0555	3.56**
u_{14}	-0.2935	0.0555	- 5.29**
u_{L3}	-0.0676	0.0680	- 0.99
u_{Q3}	0.0749	0.0392	1.91
u_{L4}	-0.0764	0.0680	- 1.12
u_{Q4}	-0.0455	0.0392	- 1.16
u_{34}	-0.1472	0.0555	- 2.65**
u_{1L3}	0.1266	0.0680	1.86
u_{1Q3}	-0.0236	0.0392	- 0.60
u_{1L4}	0.1132	0.0680	1.66
u_{1Q4}	-0.0074	0.0392	- 0.19
u_{134}	-0.0295	0.0555	- 0.53
u_{L34}	0.1080	0.0680	1.59
u_{Q34}	-0.0022	0.0392	- 0.06
u_{1L34}	-0.2279	0.0680	- 3.35**
u_{1Q34}	0.0280	0.0392	0.71

L = linear constraint for variable 2, Q = quadratic constraint for variable 2

*$P < 0.05$

**$P < 0.01$

Since four-factor effects are very difficult to interpret we ignore the significance of this effect. Hence, our "base" model is given by

$$[12]\,[13]\,[14]\,[34] \qquad (4.1)$$

which is our "best" model.

5. INTERPRETATION OF THE EXAMPLE

Each model selection method gave the same "best" model:
[12] [13] [14] [34] ($G^2 = 15.88$ on 13 d.f., $P > 0.05$)
Parameter estimates are given in Table 5 while fitted values are given in Table 6.

Table 5 Parameter estimates for the "best" model.

Parameter	Estimate	S.E.
u	3.922	0.105
$u_{1(2)}$	-1.359	0.222
$u_{2(2)}$	0.467	0.107
$u_{2(3)}$	-0.774	0.150
$u_{3(2)}$	-0.428	0.124
$u_{4(2)}$	-0.167	0.116
$u_{12(22)}$	0.065	0.233
$u_{12(23)}$	0.660	0.282
$u_{13(22)}$	0.734	0.206
$u_{14(22)}$	-1.084	0.252
$u_{34(22)}$	-0.631	0.195

Table 6 Fitted values for the "best" model for data of Table 1.

Breed	Horn length pre-dehorning	Dehorning Instrument			
		Scoop		Hodge	
		Regrowth			
		Nil	Some	Nil	Some
Brahman	≤ 2.5 cm	50.5	42.7	32.9	14.8
	2.5-3.5 cm	80.6	68.2	52.5	23.7
	≥ 3.5 cm	23.3	19.7	15.2	6.8
Sahiwal	≤ 2.5 cm	13.0	3.7	17.6	2.7
	2.5-3.5 cm	22.1	6.3	30.1	4.6
	≥ 3.5 cm	11.6	3.3	15.7	2.4

From the model breed is related to horn length pre-dehorning, dehorning instrument and regrowth while dehorning instrument and regrowth are also related. All these relationships are independent of any third variable, that is, there are no three-factor effects.

The models fitted thus far have considered all four variables as response variables. In fact a more logical structure is one where breed, horn length pre-dehorning and dehorning instrument are design variables with regrowth a response variable. Models with this structure necessarily have the terms given by [123] included. This will be discussed in more detail under Logit Models.

6. SELECTION OF A METHOD

The model selection methods of the previous sections are not the only ones available. In particular none of the methods made use of the fact that some models have direct cell estimates which could reduce computing time. This class of model was not discussed since the method of GLIM (iterative weighted least squares) takes no account of the fact.

Although each of the model selection techniques considered so far gave the same "best" model for the data of Table 1 this may not be the case for other data sets. Where there are different "best" models for the various selection methods the researcher should consider them all and determine which one (or ones) are the most useful for his situation.

It is necessary to have some criteria for assessing a particular method. Table 7 gives the number of models that had to be fitted to get to the "best" model using each method.

Table 7 Number of models to get "best" model for each method

Method	Number of Models
Brown's	23
Forward Selection	23
Backward Elimination	18
Standardized Parameter Estimates	2

Obviously Brown's method and the stepwise selection methods require a large amount of computation in comparison with the method of standardized parameter estimates. Hence where computing costs are significant their use would not be advised as a routine way of selecting a model.

Of course all these model selection techniques must be considered as merely aids to selection of a final model for a particular data set. The researcher understands his data better than any computer ever will. Hence, his input is uppermost in the selection of a model.

7. REFERENCES

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975) - Discrete Multivariate Analysis: Theory and Practice. Cambridge, Mass., The MIT Press.

Brown, M. B. (1976) - Screening effects in multidimensional contingency tables. Applied Statistics. 25:37-46.

Fienberg, S. E. (1977) - The Analysis of Cross-Classified Categorical Data. Cambridge, Mass., The MIT Press.

Goodman, L. A. (1971) - The analysis of multidimensional contingency tables: stepwise procedures and direct estimation methods for building models for multiple classifications. Technometrics. 13:33-61.