

Ideas on Variable Selection and Alternative Links in Procedure CATMOD

Kimberly Hughes DeJarnatt, John Brown University, Siloam Springs, AR

James E. Dunn, University of Arkansas, Fayetteville, AR

ABSTRACT

Two practical features are lacking in SAS procedure CATMOD: (1) for nominal categories of response, only use of baseline logit link functions admits to maximum likelihood estimation, and (2) no provision exists for automated "SELECTION=" model-building. This paper illustrates how both might be implemented, based in the first case on a class of generalized additive models previously presented to SUGI, and improvising on the basis of sweep operations in the second. Agresti-level data sets are used to illustrate the methodology. Even in those cases, we demonstrate that either model fit or power may improve by choosing alternatives to baseline logits. Considerably simplified models often result by zeroing parameters in addition to those suggested by the ANOVA table. The implementing SAS/IML code is available on request. Statisticians and other scientists faced with modeling multinomial response will find this paper of interest.

Key words: multinomial, IML, sweep, Wald statistics, logit.

INTRODUCTION

This is a story of what a next generation version of SAS procedure CATMOD might include. CATMOD's origins are in the work by Grizzle, Starnier and Koch (1969), the 'GSK model', having passed through the now defunct procedure FUNCAT in order to arrive at its current form. Fundamentally, CATMOD operates by fitting a multivariate generalized linear model (MGLM) to a link-transformed multinomial response. The baseline logit, which is appropriate for nominal responses, is the pre-programmed default link function, and adjacent category and cumulative logits are pre-programmed for ordinal responses. While additional link functions can be defined using a RESPONSE statement, only use of baseline logits admits to maximum likelihood estimation. All others default to use of generalized least squares (GLS), an artifact of the original GSK approach. In that GLS is not fully efficient at best, and non-applicable for independent multinomial samples without replication, an alternative to baseline logits seldom is considered as an option. Since Dunn (1985) pointed out the existence of alternatives to baseline logits, even for nominal responses, CATMOD of the future would make fully efficient estimation possible for these options. We demonstrate here how this might be accomplished, and present analyses of familiar data sets where improved model fits were attained.

Automated variable selection, e.g., the SELECTION = option in procedure REG, has a long history in multiple regression, and recently has been implemented for both logistic regression and proportional hazards models in procedures LOGISTIC and PHREG, respectively. Any active data analyst knows the value of these model-building algorithms. However, the problem is even more complex in the case of data sets suitable for CATMOD. Not only are the effects listed in the analysis of variance table subject to retention or deletion, but also these same effects within the individual link functions need be examined. Without limiting ourselves to baseline logits, we show that this is easily implemented using the SWEEP operator contained in the IML procedure, and again demonstrate its effectiveness using familiar data sets.

MATHEMATICAL FORMULATION

We suppose independent multinomial sampling from each of m , d -category multinomial populations, where at most a nominal relation exists among the multinomial categories. Associated with each sample are associated values of c explanatory variables, $\mathbf{x} = (x_1, \dots, x_c)'$ which may be continuous, or indicators for levels of a categorical variable, or a mixture of both. The resulting data has the form

Population	Category				Sample Size	$\mathbf{x}'$
	1	2	...	d		
1	n_{11}	n_{12}	...	$n_{1,d}$	n_{1+}	$\mathbf{x}'_1$
2	n_{21}	n_{22}	...	$n_{2,d}$	n_{2+}	$\mathbf{x}'_2$
.....						
m	n_{m1}	n_{m2}	...	$n_{m,d}$	n_{m+}	$\mathbf{x}'_m$

where $P[n_{11}, \dots, n_{1d}] = n_{1+}! \prod_{j=1}^d \pi_{1j}^{n_{1j}} / \prod_{j=1}^d n_{1j}!$, $0 < \pi_{ij} < 1$, and $\sum_{j=1}^d \pi_{ij} = 1$

for $i = 1, \dots, m$. Defining $\pi_i = [\pi_{i1}, \dots, \pi_{id}]'$, we postulate that π_i is related to $\mathbf{x}_i$ through a set of multivariate link functions, $f(\pi_i) = [f_1(\pi_i), \dots, f_u(\pi_i)]'$, where for a MGLM,

$$f(\pi_i) = (I_{(u)} \otimes \mathbf{x}_i') \beta \text{ with } \beta = (\beta_1', \dots, \beta_u')' \quad (1)$$

for $i = 1, \dots, m$.

Since necessarily $u = d - 1$ for invertibility (Dunn, 1985), we shall assume that this restriction on u always holds in the following development. A broad class of multivariate link functions (Dunn, 1985) applicable for nominal categories of response are the *generalized additive logits* (GAL), where for any monotone h , mapping $(0,1)$ into the positive real axis, $f_{ij} =$

$\ln[h(\pi_{ij})/h(\pi_{id})]$ for $j = 1, \dots, d - 1$. Its inverse is

$$\pi_{ij} = h^{-1} [h(\pi_{id}) e^{f_{ij}}], \quad (2)$$

where π_{id} solves $\sum_{j=1}^{d-1} h^{-1} [h(\pi_{id}) e^{f_{ij}}] + \pi_{id} = 1$.

Baseline logits, or *additive logits* (AL) in Aitchison's terminology (1982), with $h(\pi) = \pi$ provide a familiar example, while *additive log-logits* (ALL) and *additive tangent-logits* (ATL) with $h(\pi) = -\ln(\pi)$ and $h(\pi) = \tan(3.14159 \dots \pi/2)$, respectively, provide immediate extensions. Dunn (1985) also noted that if $\ln$ in GAL were replaced by any continuous $g: \mathbb{R}^+ \leftrightarrow \mathbb{R}$ (and $\exp$ by g^{-1}), these links also would be invertible and the class described as *generalized additive* (GA) link functions. While the examples presented here involve only the GAL class, the general approach to maximum likelihood, as well as variable selection, are applicable to the broader class of GA links.

EQUATIONS WHICH ALWAYS APPLY FOR MAXIMUM LIKELIHOOD ESTIMATION

Starting with the likelihood function $L = C \prod_{i=1}^m \prod_{j=1}^d \pi_{ij}^{n_{ij}}$,

$$\ln(L) = \sum_{i=1}^m \left[\sum_{j=1}^{d-1} n_{ij} \ln(\pi_{ij}) + n_{id} \ln(1 - \sum_{k=1}^{d-1} \pi_{ik}) \right] + \ln(C),$$

from which

$$\begin{aligned} \frac{\partial \ln(L)}{\partial \beta} &= \sum_{i=1}^m \left[\sum_{j=1}^{d-1} \frac{n_{ij}}{\pi_{ij}} \frac{\partial \pi_{ij}}{\partial \beta} - \frac{n_{id}}{1 - \sum_{k=1}^{d-1} \pi_{ik}} \sum_{k=1}^{d-1} \frac{\partial \pi_{ik}}{\partial \beta} \right] \\ &= \sum_{i=1}^m \left[\sum_{j=1}^{d-1} \left(\frac{n_{ij}}{\pi_{ij}} - \frac{n_{id}}{\pi_{id}} \right) \frac{\partial \pi_{ij}}{\partial \beta} \right] = \mathbf{w}(\beta), \end{aligned} \quad (3)$$

defining the maximum likelihood estimator (MLE) to be $\hat{\beta}$

satisfying $\mathbf{w}(\hat{\beta}) = \mathbf{0}$. In order to implement Fisher's method of scoring, the matrix of second derivatives of log-likelihood is given by $\partial \mathbf{w} / \partial \beta =$

$$\sum_{i=1}^m \left[\sum_{j=1}^{d-1} \left\{ -\frac{n_{ij}}{\pi_{ij}^2} \frac{\partial \pi_{ij}}{\partial \beta} - \frac{n_{id}}{\pi_{id}^2} \sum_{k=1}^{d-1} \frac{\partial \pi_{ik}}{\partial \beta} \right\} \left(\frac{\partial \pi_{ij}}{\partial \beta} \right) + \sum_{j=1}^{d-1} \left\{ \frac{n_{ij}}{\pi_{ij}} - \frac{n_{id}}{\pi_{id}} \right\} \frac{\partial}{\partial \beta} \left(\frac{\partial \pi_{ij}}{\partial \beta} \right) \right].$$

Replacing n_{ij} by its expectation, $E(n_{ij}) = n_{i\cdot} \pi_{ij}$ yields

$$E(\partial \mathbf{w} / \partial \beta) = -\mathbf{W} =$$

$$-\sum_{i=1}^m n_{i\cdot} \left[\sum_{j=1}^{d-1} \frac{1}{\pi_{ij}} \frac{\partial \pi_{ij}}{\partial \beta} \left(\frac{\partial \pi_{ij}}{\partial \beta} \right) + \frac{1}{\pi_{id}} \sum_{k=1}^{d-1} \frac{\partial \pi_{ik}}{\partial \beta} \left(\sum_{j=1}^{d-1} \frac{\partial \pi_{ij}}{\partial \beta} \right) \right], \quad (4)$$

a symmetric matrix, where $\mathbf{V} = \mathbf{W}^{-1}$ is the asymptotic covariance matrix of $\hat{\beta}$. Solution is by means of a Newton-Raphson iteration, $\hat{\beta}_{s+1} = \hat{\beta}_s + \mathbf{W}(\hat{\beta}_s)^{-1} \mathbf{w}(\hat{\beta}_s)$, or more concisely, $\hat{\beta}_{s+1}$ solves the linear equations $\mathbf{W}(\hat{\beta}_s) \hat{\beta}_{s+1} = \mathbf{w}(\hat{\beta}_s)$, where the right hand side $\mathbf{w}(\hat{\beta}_s) = [\mathbf{W}(\hat{\beta}_s) \hat{\beta}_s + \mathbf{w}(\hat{\beta}_s)]$, is a vector of "working values" at iteration step s as traditionally defined, e.g., Finney (1971).

From equations (3) and (4) we see that $\hat{\beta}$ depends on choice of link functions only through $\partial \pi_{ij} / \partial \beta$. But even this has a general form for GAL and GA links since from equation (2), $\pi_{ij} = \pi_{ij}(f_{ij}, \pi_{id})$ for $j = 1, \dots, d-1$. Thus,

$$\begin{aligned} \frac{\partial \pi_{ij}}{\partial \beta} &= \begin{bmatrix} \frac{\partial \pi_{ij}}{\partial \beta_1} \\ \vdots \\ \frac{\partial \pi_{ij}}{\partial \beta_{d-1}} \end{bmatrix} = \begin{bmatrix} \delta_{j1} \frac{\partial \pi_{ij}}{\partial f_{ij}} \frac{\partial f_{ij}}{\partial \beta_1} + \frac{\partial \pi_{ij}}{\partial \pi_{id}} \frac{\partial \pi_{id}}{\partial \beta_1} \\ \vdots \\ \delta_{jd-1} \frac{\partial \pi_{ij}}{\partial f_{i,d-1}} \frac{\partial f_{i,d-1}}{\partial \beta_{d-1}} + \frac{\partial \pi_{ij}}{\partial \pi_{id}} \frac{\partial \pi_{id}}{\partial \beta_{d-1}} \end{bmatrix} \\ &= \begin{bmatrix} \delta_{j1} \frac{\partial \pi_{ij}}{\partial f_{ij}} \mathbf{x}_i - \frac{\partial \pi_{ij}}{\partial \pi_{id}} \sum_{k=1}^{d-1} \frac{\partial \pi_{ik}}{\partial \beta_1} \\ \vdots \\ \delta_{jd-1} \frac{\partial \pi_{ij}}{\partial f_{i,d-1}} \mathbf{x}_i - \frac{\partial \pi_{ij}}{\partial \pi_{id}} \sum_{k=1}^{d-1} \frac{\partial \pi_{ik}}{\partial \beta_{d-1}} \end{bmatrix}, \end{aligned} \quad (5)$$

where δ_{rs} is Kronecker's delta, equal to 1 if $r = s$ and 0 otherwise. Equation (5) can be simplified. Summing over $j = 1, \dots, d-1$, obtain

$$\sum_{j=1}^{d-1} \frac{\partial \pi_{ij}}{\partial \beta} = \begin{bmatrix} \sum_{j=1}^{d-1} \frac{\partial \pi_{ij}}{\partial \beta_1} \\ \vdots \\ \sum_{j=1}^{d-1} \frac{\partial \pi_{ij}}{\partial \beta_{d-1}} \end{bmatrix} = \begin{bmatrix} \frac{\partial \pi_{i1}}{\partial f_{i1}} \mathbf{x}_i - \sum_{j=1}^{d-1} \frac{\partial \pi_{ij}}{\partial \pi_{id}} \sum_{k=1}^{d-1} \frac{\partial \pi_{ik}}{\partial \beta_1} \\ \vdots \\ \frac{\partial \pi_{i,d-1}}{\partial f_{i,d-1}} \mathbf{x}_i - \sum_{j=1}^{d-1} \frac{\partial \pi_{ij}}{\partial \pi_{id}} \sum_{k=1}^{d-1} \frac{\partial \pi_{ik}}{\partial \beta_{d-1}} \end{bmatrix}$$

which shares the common term $\sum_{j=1}^{d-1} \frac{\partial \pi_{ij}}{\partial \beta}$ on both sides. Solving for this yields

$$\sum_{j=1}^{d-1} \frac{\partial \pi_{ij}}{\partial \beta} = \left(1 + \sum_{k=1}^{d-1} \frac{\partial \pi_{ik}}{\partial \pi_{id}} \right)^{-1} \begin{bmatrix} \frac{\partial \pi_{i1}}{\partial f_{i1}} \mathbf{x}_i \\ \vdots \\ \frac{\partial \pi_{i,d-1}}{\partial f_{i,d-1}} \mathbf{x}_i \end{bmatrix}, \quad (6)$$

which when substituted in equation (5) yields

$$\frac{\partial \pi_{ij}}{\partial \beta} = \begin{bmatrix} [\delta_{j1} - (1 + \sum_{k=1}^{d-1} \frac{\partial \pi_{ik}}{\partial \pi_{id}})^{-1} \frac{\partial \pi_{ij}}{\partial \pi_{id}}] \frac{\partial \pi_{i1}}{\partial f_{i1}} \mathbf{x}_i \\ \vdots \\ [\delta_{jd-1} - (1 + \sum_{k=1}^{d-1} \frac{\partial \pi_{ik}}{\partial \pi_{id}})^{-1} \frac{\partial \pi_{ij}}{\partial \pi_{id}}] \frac{\partial \pi_{i,d-1}}{\partial f_{i,d-1}} \mathbf{x}_i \end{bmatrix}. \quad (7)$$

Equations (6) and (7) are readily computable as intermediate steps in evaluating equations (3) and (4).

ANOVA TESTS USING SWEEP OPERATIONS

Since sweeping $[\mathbf{W}(\hat{\beta}_s) : \mathbf{w}(\hat{\beta}_s)]$ with respect to all rows at each iterative step yields $[\mathbf{W}(\hat{\beta}_s)^{-1} : \hat{\beta}_{s+1}]$, at the

MLE solution, $\hat{\beta}$, $\text{var}[\hat{\beta}]$ is estimated by $\mathbf{W}(\hat{\beta}_s)^{-1}$. Once convergence is attained, then a generalized Wald statistic defined by $X^2 = \hat{\beta}' \mathbf{G}' [\mathbf{G} \mathbf{W}(\hat{\beta}_s)^{-1} \mathbf{G}']^{-1} \mathbf{G} \hat{\beta}$ for testing $H_0 : \mathbf{G} \hat{\beta} = \mathbf{0}$, where $\mathbf{G}$ is any column permutation of $[\mathbf{I}_{(m)} : \mathbf{0}]$, is obtained by sweeping

$$\begin{bmatrix} \mathbf{W}(\hat{\beta}_s)^{-1} & \hat{\beta} \\ \hat{\beta}' & 0 \end{bmatrix} \quad (8)$$

with respect to m rows corresponding to subscripts of non-zero columns of $\mathbf{G}$. The result of the sweep is $\begin{bmatrix} \bullet & \bullet \\ \bullet & -X^2 \end{bmatrix}$, so that one

only needs to change the sign of the bottom corner element to obtain the test statistic, $X^2 \sim \chi_{(m)}^2$ under H_0 . Since matrix (8) is unaffected by the sweep, it may be swept again to test other hypotheses.

Consider an example given by Agresti (1996) in which alligators from five lakes and two size classes were classified with respect to their primary food choice. The data appears in Table 1.

Table 1. Alligators classified by primary food choice.

Lake	Primary Food Choice					
	Size (m)	Fish	Invert.	Reptile	Bird	Other
Hancock	≤ 2.3	23	4	2	2	8
	> 2.3	7	0	1	3	5
Oklawaha	≤ 2.3	5	11	1	0	3
	> 2.3	13	8	6	1	0
Trafford	≤ 2.3	5	11	2	1	5
	> 2.3	8	7	6	3	5
George	≤ 2.3	16	19	1	2	3
	> 2.3	17	1	0	1	3

STATISTICS AND DATA ANALYSIS

Table 2 shows identical chi-squared statistics given by CATMOD using default AL links and those resulting from the sweep operations described here. Since the explanatory variables, Lake and Size Class, are categorical, zero-sum restrictions were imposed (as does CATMOD) before the sweep operations. Because the order of parameters corresponds to that of equation (1) (unlike CATMOD), rows 1, 6, 11, and 16 of matrix (8) were swept to obtain SS for Intercept; rows 2, 3, 4, 7, 8, 9, 12, 13, 14, 17, 18, and 19 were swept to obtain SS for Lakes; and rows 5, 10, 15, and 20 were swept to obtain SS for Size Class. This was repeated using ALL links, with results also shown in Table 2.

Table 2. Analysis of variance of primary food choice data in Table 1 using additive logistic (AL) and additive log-logistic (ALL) links.

Source	d.f.	AL X^2	p-value	ALL X^2	p-value
Intercept	4	70.39	<0.0001	84.02	<0.0001
Lake	12	35.49	3.91E-4	48.30	2.78E-6
Size	4	18.76	8.76E-4	20.29	4.37E-4
Fit	12	17.08	0.1466	17.85	0.1204

Even though lack-of-fit X^2 was marginally smaller for AL links, use of ALL links resulted in an increase of 12.81 in Lake X^2 . Since the degrees of freedom are identical in both cases, the ALL-based test for Lake effects has increased power over that based on AL links. Since our focus until now primarily had been on comparing lack-of-fit under alternative choices of links, the potential for increased power as a result of link choice remains unexplored territory.

In fact it is not difficult to find examples in the published literature where use of ALL links provides a better fit than use of AL links. Grizzle, *et al.* (1969) illustrated GSK methodology using numbers of depletions (deaths) incurred in litters of mice classified by litter size and exposure to either treatments A or B. The data appears in Table 3, and the resulting ANOVA, treating litter size as quantitative, is shown in Table 4. Use of ALL links rather than AL reduced lack-of-fit X^2 from 10.21 to 9.04. Note that a marked increase in the X^2 statistic for litter size also has occurred.

Table 3. Number of litters showing 0, 1 or 2+ depletions.

Litter Size	Treatment	Number of depletions		
		0	1	2+
7	A	58	11	5
	B	75	19	7
8	A	49	14	10
	B	58	17	8
9	A	33	18	15
	B	45	22	10
10	A	15	13	15
	B	39	22	18
11	A	4	12	17
	B	5	15	8

Table 4. Analysis of variance of number of mice depletions treating litter size as quantitative using additive logit (AL) and additive log-logit (ALL) links.

Source	d.f.	AL X^2	p-value	ALL X^2	p-value
Intercept	2	107.15	<0.0001	123.03	<0.0001
Size	2	78.41	<0.0001	95.81	<0.0001
Treatment	2	6.71	0.0350	5.68	0.0585
Fit	14	10.21	0.7464	9.04	0.8287

BACKWARD ELIMINATION USING SWEEP

Quite often, even after eliminating factors and interactions based on the ANOVA table, many of the 1 d.f. Wald statistics, i.e., $X^2 = (\text{estimate/standard error})^2 \sim \chi^2_{(1)}$, are not significantly different from zero, thus suggesting that a simpler model may be attained. The relative magnitudes of these, from small to large, suggest a tentative order for elimination. Since conditional odds ratios ala Agresti (1996) are identically 1 for all those non-intercept parameters which are zeroed, this often leads to considerable economy of interpretation. In order to eliminate additional parameters, one simply does successive sweeps of matrix (8) with respect to rows corresponding to indices of increasing Wald statistics until lack of statistical significance of the entire set of zeroed parameters is no longer obtained. This is illustrated in Table 5, based on the primary food choice data in Table 1, and using both AL and ALL links. Profiles are in order: bird, fish, invertebrates, other, reptiles (using reptiles as the baseline); George, Hancock, Oklawaha, Trafford; and >2.3, ≤ 2.3. Choosing $\alpha = 0.05$ as the nominal significance level, selection stopped for AL links with 10 parameters declared not significantly different from zero ($p = 0.105$), while for ALL links, selection stopped with 7 parameters declared not significantly different from zero ($p = 0.071$). This difference seems attributable to the fact that Wald statistics generally were larger using ALL rather than AL links, which corresponds to a similar trend already noted in the ANOVA statistics in Table 2. With the exception of parameter #12, all other parameters selected using ALL links were included in the set selected using AL links.

Table 5. Order of selection of parameters to be zeroed in models for alligator primary food choice, using additive logits (AL) and additive log-logits (ALL).

Index	Parameters	AL X^2	Order	ALL X^2	Order
1	β_1	0.388	4	0.864	3
2		2.149	10	3.053	
3		1.349	8	1.603	7
4		3.756		4.306	
5		0.120	2	0.057	1
6	β_2	29.03		52.32	
7		4.371		6.881	
8		0.555	5	3.488	
9		2.882		2.698	
10		0.367	3	0.232	2
11	β_3	9.109		14.38	
12		3.676		2.494	8
13		4.084		8.979	
14		0.006	1	1.476	6
15		9.008		13.39	
16	β_4	3.168		3.155	
17		1.574	9	0.942	4
18		1.166	7	2.731	
19		4.837		4.947	
20		1.099	6	1.450	5

As a final remark in this section, note that we have been able to produce Table 5 based on only a single model fit for each set of links. This is a distinct advantage of backward elimination since model-fitting for MGLM's tends to be computationally expensive in that the size of parameter sets associated with MGLM's tends to be explosive. We have not based our model-building decisions entirely on 1 d.f. Wald statistics, but rather do a single, multiple d.f. test to make a decision about whether or not to zero the entire tentative subset of parameters. It's a simple computational step using sweep operations, and leads to a single p-value at the end.

CONCLUSION

For categorical predictors, the preliminary step of reparameterizing to full rank, e.g., using zero-sum restrictions, has not been essential in any of the data sets we've examined, since all of the computational steps basically are identical regardless of whether a true or a g^2 -inverse is used. This increasingly has become an encumbrance for CATMOD users who have forgotten or never learned the concept of reparameterizing a model to full rank.

For the alligator primary food choice data, a general increase in the size of Wald statistics was observed, both individually and in the ANOVA table, as a result of replacing AL links with ALL links. Selection of 7 parameters to zero in the latter case, compared to 10 using AL links is associated with this increase. Clearly this is attributable to standard errors using ALL links which are half or less than those based on AL links. We have not yet explored whether this is a general occurrence, or simply specific to this data set.

For some non-trivial data sets encountered in our consulting work, we have been able to obtain major model simplifications using an *ad hoc* approach to parameter selection now carried on systematically by our backward elimination algorithm. In the case of the alligator primary food choice example, we would focus on interpreting 7 or perhaps 9 conditional odds ratios involving reptiles, rather than the initial undifferentiated 16, 9 or 7 of which we expect to be close to 1. We are aware, however, that one cannot always fit a "full model" to use as a starting point for backward elimination. We can, however, compute efficient score statistics without fitting the model, and this is an avenue which we are pursuing.

REFERENCES

- Agresti, A. 1996. *An Introduction to Categorical Data Analysis*, John Wiley & Sons, New York.
- Aitchison, J. 1982. The statistical analysis of compositional data. *J. of the Royal Statistical Society (B)* 44:139 – 177.
- Dunn, J.E. 1985. The role of invertibility and symmetry of polytomous metameters for GSK and other related analyses. *Proc. of 10th Annual Conf. of SAS Users Group*, pp. 990 – 999.
- Finney, D.J. 1971. *Probit Analysis*, 3rd edition, Cambridge University Press.
- Grizzle, J.E., Stamer, C.F., and Koch, G.G. 1969. Analysis of categorical data by linear models. *Biometrics* 25:489 – 504.

CONTACT INFORMATION

Your comments and questions are valued and encouraged.
Contact the author at:

Kimberly Hughes DeJarnatt
John Brown University
2000 W. University
Siloam Springs, AR 72762
(501) 524-7279
khughes@jbu.edu