Computational Statistics & Data Analysis 2 (1984) 159–165
North-Holland

The choice of a log-linear model using a C_p-type statistic

E. Tejumola JOLAYEMI
Department of Mathematics, Ahmadu Bello University, Zaria, Nigeria

Morton B. BROWN
Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

Received March 1984
Revised April 1984

Abstract: Most methods of selecting an appropriate log-linear model for categorical data are sensitive to the underlying distributional assumptions. However, there are many situations in which the assumption that the data are randomly chosen from an underlying Poisson, multinomial or product-multinomial distribution cannot be sustained. In these cases we propose a criterion to select among log-linear models that is an analogue of the C_p statistic for regression models and describe a method to estimate the denominator of this statistic.

Keywords: Categorical data, Log-linear models, Model selection, C_p.

1. Introduction

Associations between categorical variables are often studied by fitting log-linear models to data summarized as a multidimensional contingency table. Goodman [5,6], Brown [2], Wermuth [12] and others have proposed methods to select the 'best' log-linear model in the absence of a priori knowledge of the underlying relationships between the variables. In all these methods, the choice of the 'best' model involves chi-square tests of significance of the differences in the lack-of-fit statistics between pairs of hierarchical models. The asymptotic distribution theory of the chi-square tests is based on the assumption that the data in the contingency table are a result of independent random sampling from a Poisson, multinomial or product multinomial distribution

In many applications the assumption either of independence or of the underlying distributional form is inappropriate. When these assumptions are violated, the variability in the data is often increased. Williams [13], Coffey [3] and others fit logistic linear models to data containing extra-binomial variation. However, these approaches require the parametric modelling of the underlying distributional form.

An alternative approach is to approximate the distribution of the sample proportions by one in which the variance is proportional to the mean. The log-linear model is an example of a quasi-likelihood function [10,11] that can be fitted to the data using weighted least squares without the need to specify the entire likelihood. Formal tests of significance to select a final model may then not be appropriate. The investigator may still want to select the 'best' model(s) using a heuristic criterion.

A similar problem is that of variable selection in multiple linear regression. Mallows [9] proposed the C_p statistic to compare alternative regression models. Minimizing the C_p statistics is often used as a criterion to select the 'best' subset regression [7]. However, the C_p statistic does not provide a formal mechanism for hypothesis testing.

We extend the C_p method to the selection of the 'best' log-linear model when the variances of the proportions can be assumed to be proportional to the proportions. Use of this C_p analogue enables models with different numbers of parameters to be compared without relying on the distribution of the chi-square statistic. We describe how to estimate the denominator of this C_p analogue.

We conclude with an example based on census data. As is common in such sets, all chi-square statistics are very large. The C_p method provides a rationale to select an unsaturated model to describe the relationships between the factors.

2. The C_p statistic

The multiple linear regression model can be written as

$$E(y) = X\beta$$

where y is an $(n \times 1)$ column vector of observed frequencies, $E(y)$ is the expectation of y, X is an $(n \times p)$ design matrix, and β is a $(p \times 1)$ column vector of parameters.

For a multiple linear regression model Mallows' C_p is defined as

$$C_p = \frac{\mathrm{SSE}_p}{\hat{\sigma}^2} - (n - 2p)$$

where n is the number of observations, p is the number of parameters estimated in the model, SSE_p the sum of squares of error (residuals) for this model, and $\hat{\sigma}^2$ the estimate of the variance (usually computed as the mean square error when all variables are entered into the regression).

If the model being fitted is appropriate, the expected value of SSE_p is $(n-p)\sigma^2$ and C_p is of magnitude p. The addition of a new variable to the model reduces C_p only if the sum of squares of error decreases by at least twice the estimate of the variance (i.e., $2\hat{\sigma}^2$).

A modification of the C_p statistic is to replace $(n - 2p)$ by $(n - kp)$ where the value of k, called the penalty, is specified by the investigator. Then, the addition of a new variable to the model reduces C_p only when the decrease in the error sum of squares exceeds $k\hat{\sigma}^2$. In this article we consider the case when k is two (the original C_p) but only minor changes are required to allow other values of k.

In an analogous manner the log-linear model can be written as

$$\ln \boldsymbol{F} = \ln E(\boldsymbol{f}) = X\boldsymbol{\beta}$$

where $\boldsymbol{f}$ is an $n \times 1$ vector of observed cell frequencies, $\boldsymbol{F}$ is a vector of expected cell frequencies, and X and $\boldsymbol{\beta}$ are as defined above in the regression model.

Under independent Poisson, multinomial or product multinomial sampling, the maximum likelihood solution can be obtained by a Fisher scoring technique [8] that solves iteratively for the parameter vector $\boldsymbol{\beta}$ by finding a solution (subject to constraints) to the normal equations of

$$(\boldsymbol{f} - \boldsymbol{F})' W^{-} (\boldsymbol{f} - \boldsymbol{F})$$

where W^{-} is a generalized inverse of the variance–covariance matrix of $\boldsymbol{f}$.

The likelihood ratio statistic is

$$G^2 = -2 \sum f_i \ln(f_i / F_i)$$

where the sum is over all cells. Under the distributional assumptions and when the model is appropriate, G^2 is asymptotically distributed as a chi-square statistic with $n - p$ degrees of freedom (df) and, hence, its expected value is $n - p$. The statistic G^2 is also asymptotically equivalent to the Pearson chi-square statistic

$$X^2 = \sum (f_i - F_1)^2 / F_i.$$

Violations of the distributional assumptions are likely to affect the expectation of G^2. Finney [4], for example, used a 'heterogeneity factor', estimated by $X^2/(n-p)$, to rescale tests-of-fit. When the test-of-fit of a log-linear model that should fit the data appropriately yields a value of $G^2/(n-p)$ that is not near unity, it may be possible to attribute the excess to violations of the distributional assumptions rather than to an inappropriate specification of the parameters of the log-linear model.

In order to use the C_p analogue, it is sufficient to assume that the variances of the sample proportions are proportional to their expected values. The proposed C_p analogue is

$$C_p = \frac{G^2}{\hat{\sigma}^2} - (n - 2p)$$

where G^2 is the usual test-of-fit statistic for the model of interest and

$$\hat{\sigma}^2 = G_b^2 / f_b$$

where b represents the base model whose fit theoretically should be adequate, G_b^2 and f_b are the likelihood ratio chi-square statistic for the base model and its degrees of freedom, and n is the number of cells in the table. When the distributional assumptions are fulfilled, $\hat{\sigma}^2$ can be replaced by unity.

The value of C_p is reduced when a term (or terms) is added to the model if and only if the change in G^2 is at least twice the product of $\hat{\sigma}^2$ and the df associated with the term being added.

3. The choice of σ^2

In the absence of a priori knowledge about the appropriate model, the choice of the base model is important. If the base model contains too many parameters, the lack-of-fit statistic G^2 may have few df and, as a result, have large variability. If the base model is too parsimonious and omits some important interactions, G^2 will be overestimated.

The problem of estimating σ^2 has a parallel in the estimation of the error mean square in a multiway analysis of variance with one observation per cell. The usual initial estimate of the error mean square is from the mean square of the highest order interaction. However, this estimate may have few df. Therefore, an empirical rule that is often used is to pool sums of squares of other interaction terms into the error sum of squares when the mean square of the interactions are less than twice the error mean square.

In a similar manner we recommend that the initial estimate for σ^2 be

$$\hat{\sigma}^2 = G_k^2/f_k$$

where k represents the model of order k (that contains all k-factor interactions but no $(k+1)$-factor interaction) and where k is the lowest order model such that the change in G^2 between two successive models divided by the change in df is less than $2G_{k+1}^2/f_{k+1}$ for all models of higher order. Chi-square tests of marginal and partial association [1,2] are then computed for each interaction. Using the above initial estimate of $\hat{\sigma}^2$, any interaction that has either a test of partial association or a test of marginal association that exceeds $2\hat{\sigma}^2$ is included in the model that is to be fitted at the first step. At each step $\hat{\sigma}^2$ is recalculated from the lack-of-fit of the current model. The iterative procedure steps when no interaction term can be added to or removed from the model. For this final model

$$C_p = \frac{G^2}{\hat{\sigma}^2} - (n-2p) = (n-p) - (n-2p) = p,$$

since $\hat{\sigma}^2 = G^2/(n-p)$.

4. The Kenya 1969 census data

Table 1 reports a subset of data from the Kenya 1969 census. The data are cross-classified according to Age (12 categories), Sex, Father's status (alive or dead) and Mother's status. The cell counts are in thousands.

Since the data are from a census, one may question the assumption that there is a superpopulation from which a random sample was taken. Also, since the frequency counts are so large, any hypothesis that is tested will almost invariably be rejected. Therefore, the goal of fitting a log-linear model is restricted to defining a parsimonious model that includes the 'important' interactions. One way of choosing such a parsimonious model is by use of the C_p analogue.

In the first panel of Table 2 are the simultaneous tests of different orders. The

Table 1
Kenya 1969 census data (in thousands)

Year	Males					Females				
	F, M	F, −M	−F, M	−F, −M	Total	F, M	F, −M	−F, M	−F, −M	Total
0–4	995	5	44	3	1047	983	5	45	3	1036
5–9	831	14	57	4	906	808	13	53	4	878
10–14	614	20	67	7	708	584	16	60	6	666
15–19	438	25	78	14	555	427	24	76	12	539
20–24	287	28	86	22	423	299	31	90	26	446
25–29	190	28	90	33	341	220	36	102	46	404
30–34	119	26	85	48	278	119	29	86	62	296
35–39	80	23	80	64	247	79	25	78	76	258
40–49	66	28	106	158	358	61	27	94	177	359
50–59	16	10	48	172	246	16	9	36	177	238
60–69	6	4	15	150	175	7	2	10	136	155
70+	4	1	4	121	130	6	1	3	113	123
Total	3646	212	760	796	5414	3609	218	733	838	5398

M Mother alive; −M Mother not alive; F Father alive; −F Father not alive.

initial estimate of the variance based on the highest-order interaction is $\hat{\sigma}^2 = 1470/11 = 133.64$. In the second panel of Table 2 are tests of partial and marginal association. These results should be compared to $2\hat{\sigma}^2 = 267.3$. The ratios of the G^2 to their dfs for the interactions FS, AMS and AFS are less than $2\hat{\sigma}^2$. Therefore

Table 2
Simultaneous tests of all interactions with k or more factors (G^2 in thousands)

k-Factor	df	G^2
1	95	19705.62
2	81	9446.63
3	45	66.88
4	11	1.47

Tests of partial and marginal association (G^2 in thousands)

Interaction	df	Partial association G^2	Marginal association G^2
AM	11	1816.01	4425.67
AF	11	2041.55	4653.38
AS	11	12.14	10.17
MF	1	286.62	2898.52
MS	1	3.40	1.40
FS	1	0.00	0.06
AMF	11	61.20	60.95
AMS	11	1.30	2.83
AFS	11	0.81	1.29
MFS	1	0.87	0.16
AMFS	11	1.47	1.47

the hierarchical model defined by the configurations AS, AMF, MFS is fitted to the data. The revised estimate of $\hat{\sigma}^2$ is 126.06 and no interactions are added or deleted at the next step.

Models selected by criteria that are based on the significance of G^2 would, not surprisingly, choose the saturated model.

5. Discussion

The usual methods of model selection test for the inclusion or exclusion of an interaction term are based on chi-square tests. However, the distribution of the test statistic is asymptotically chi-square only if the underlying distribution from which the sample is drawn is Poisson, multinomial or product-multinomial.

Various ad hoc methods are used when the distributional assumptions are violated. An example is to choose the most parsimonious model whose test of lack-of-fit is less than a specified fraction of that of a base model. This rationale is based on the desire to explain a certain fraction of the variation – here called lack-of-fit.

The C_p method compares the magnitudes of the tests of the interactions with that of the test of lack-of-fit. Therefore, the criterion to include or exclude a term is based on its relative magnitude compared to the other terms.

Several cautions are necessary. Except in repeated-measures or case-control experiments it is unlikely that a model of smaller than expected variation is appropriate. Therefore, unless the design indicated a high likelihood of subnormal variation, we suggest that the lower bound for $\hat{\sigma}^2$ be unity. It may be noted that in many published data sets (see e.g., [12]) the test of the highest order interaction is very nonsignificant. In these data sets the use of $\hat{\sigma}^2$, rather than setting $\hat{\sigma}^2$ to unity, would include too many terms in the model.

Another method of model selection is analogous to the use of the adjusted R^2 in multiple regression. This criterion reduces to choosing the model that minimizes the ratio of the test of lack-of-fit to its df. As a result, interaction are included if their test statistic divided by df is greater than unity. Since the expected value of a central chi-square variate divided by its df is unity, the effective level of significance for using this criterion is nearer 50% than 5% even when the underlying model is multinomial or Poisson and, therefore, too many terms may be selected for the model.

The adjusted R^2 criterion and more classical methods based on chi-square tests of significance are not appropriate for the more general sampling models described here since their criteria are not modified to allow for increased variances and, therefore, they would accept too many terms into the model.

Since the appropriate model for many large data sets is the saturated model, the C_p criterion allows the investigator to separate interactions into two groups: large and small. An examination of the tests of marginal and partial association (divided by degrees of freedom) also allow judgements about magnitude. The C_p's advantage is that it provides a standard cut-off point.

References

[1] J. Benedetti and M.B. Brown, Strategies for the selection of log-linear models, *Biometrics* **34** (1978) 680–686.

[2] M.B. Brown, Screening effects in multidimensional tables, *Appl. Statist.* **25** (1976) 37–46.

[3] M. Coffey, A class of categorical models with random effects and their estimation by maximum likelihood, Presented at the *Biometric Society Spring Meeting*, Nashville, TN (March 21, 1983).

[4] D.J. Finney, *Probit Analysis* (University Printing House, Cambridge, England, 1971).

[5] L.A. Goodman, The multivariate analysis of qualitative data: interactions among multiple classifications, *J. Amer. Statist. Assoc.* **65** (1970) 226–256.

[6] L.A. Goodman, The analysis of multidimensional contingency tables: Stepwise procedures for building models for multiple classifications, *Technometrics* **13** (1971) 33–61.

[7] J.W. Gorman and R.J. Toman, Selection of variables for fitting equations to data, *Technometrics* **8** (1966) 27–51.

[8] R.I. Jennrich and R.H. Moore, Maximum likelihood estimation by means of nonlinear least squares, *Proceedings of the Statistical Computing Section, American Statistical Association* (1975) 57–65.

[9] C.L. Mallows, Some comments on C_p, *Technometrics* **15** (1973) 661–675.

[10] P. McCullagh, Quasi-likelihood functions, *Ann. Statist.* **11** (1983) 59–67.

[11] R.W.M. Wedderburn, Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method, *Biometrika* **61** (1974) 439–447.

[12] N. Wermuth, Model search among multiplicative models, *Biometrics* **32** (1976) 253–263.

[13] D.A. Williams, Extra-binomial variation in logistic linear models, *Appl. Statist.* **31** (1982) 144–148.