Numerical Examination on Sensitivity in Latent Class Analysisby Andersen's Diagnostics

森　田　築　雄

美作大学・美作大学短期大学部紀要（通巻第50号抜刷）

美作大学・美作大学短期大学部紀要
2005,　Vol. 50.　31 ～ 37

論　文

Numerical Examination on Sensitivity in Latent Class Analysis by Andersen's Diagnostics

Tsukio Morita

Abstract

Under the assumption of the local independence, latent class analysis can be reduced to a parametric multinomial distribution. For sensitivity analysis in a multinomial distribution model, in general, a needed diagnostics is the diagnostics that measures the influence by deleting a single cell, rather than an observation. In this paper we apply Andersen's diagnostics, which is suggested for a parametric multinomial distribution, to two artificial latent class models and data set analysed by Stouffer and Toby, and examine their numerical results. Also it is shown that the reliability of estimated parameters in latent class analysis is evaluated by thier asymptotic variances at ML estimates.

1 Introduction

In regression analysis special attention is devoted to identifying observations, which give a significant influence to the model fit or to the estimated regression coefficients (see Cook and Weisberg, 1982). Pregibon (1981) has discussed the sensitivity to outlying responses and extreme points for a maximum likelihood fit of a logistic regression model. Similar consideration is devoted to principle component analysis (Critchley, 1985) and factor analysis (Tanaka and Odaka, 1989) by using the influence function. For a multinomial distribution, Andersen (1992) has given diagnostics as measures of model deviation and of the influence of deleting a single cell. Andersen's diagnostics is considered as a useful tool for sensitivity analysis in latent class analysis, since by the assumption of local independence it can be reduced to a multinomial distribution. In Section 2, we summarize latent class analysis and Andersen's diagnostics. In section 3, from the viewpoints of Andersen's diagnostics and aymptotic variances, we will investigate two artificial latent class models and data set examined by Stouffer and Toby (1951), Lazersfeld and Henry (1968), and Goodman (1974a, 1975, 1979) (also see, McCutcheon, 1987; Hagenaars and McCutcheon, 2002).

2 Latent Class Analysis and Andersen's Diagnostics

Latent class analysis (LCA) is a technique for analyzing relationships in categorical data. The basic premise of the study of latent variables is that the latent variable explains the relationships between the observed variables. There are several typical methods for obtaining estimates of parameters in LCA (Green, 1951; Gibson, 1955; MacHugh, 1956, *et al.*). We can use here the maximum likelihood method in analyzing sensitivity in LCA by the assumption of local independence that the relationships observed among a set of variable are found to be zero within the categories of the latent variable. Suppose that there are n items and m latent classes and that for each item, an individual belongs to one and only one of m classes. Under the condition of local independence latent structure equation is, then, written as

$$\eta_s = \sum_{t=1}^{m} w_t \prod_{i=1}^{n} \pi_{it}^{y_i^{(s)}} (1 - \pi_{it})^{1-y_i^{(s)}}, \quad s = 1, \cdots, 2^n$$

where the latent class probability (w_t) is the probability that a randomly selected observation in the sample is located in latent class t , the conditional probability (π_{it}) is the probability that an individual of a latent class t responds positively to item i, and ($y_1^{(s)}, \ldots, y_n^{(s)}$) with $y_i^{(s)}$=0 or 1 (i=1, . . . , n) denotes the s - th response pattern (cell s), where for item i in pattern s , $y_i^{(s)}$=0 indicates the negative response and $y_i^{(s)}$=1 the positive response. Then η_s represents the probability for the s - th response pattern. For a single latent variable, we can express the restriction as $\sum_{t=1}^{m} w_t=1$.

Let $I=2^n$ and $n=(n_1, \ldots, n_I)$ be the frequency vector for I response patterns, then $n \sim Mul\,(N,\ \eta\,(\theta))$, where $N=\sum_j^I n_j$ and $\theta=(w_1, \ldots, w_{m\text{-}1},\ \pi_{11}, \ldots,\ \pi_{nm})$.

In a multinomial model the quantities of interest are the observed counts in the cell. As pointed out by Andersen (1992), the term in the likelihood function corresponding to cells are not independent, hence it does not make sense merely to remove a term in the likelihood function. Andersen (1992) derives Cook's distance by substituting the cell probabilities by the conditional probabilities given that a cell is omitted and then forming the log-likelihood function as a sum of contributions from the remaining cells, and further shows that Cook's distance can be approximated by an expression which does not need a re-estimation of the parameters. Let $D=(d_{sp})=\partial \log \eta_s / \partial \theta_p$, $I\times P$, P being $m\,(n+1)-1$ and $V=diag(N\eta_s)$, s =1, . . . , I . The two important diagnostics measures that he gives are as follows: First, the standardized residuals

$$r_s = (n_s - N\hat{\eta}_s)/\sqrt{N\hat{\eta}_s(1-\hat{\eta}_s-\hat{h}_s)}$$

where $\hat{\eta}=\eta(\hat{\theta})$, $\hat{\theta}$ being the vector of ML estimates based on all observations, and $\hat{h}_s$ is h_s evaluated at $\hat{\theta}$,

$$h_s = N\eta_s \sum_{p=1}^{P}\sum_{q=1}^{P}(\partial \log \eta_s/\partial\theta_p)(\partial \log \eta_s/\partial\theta_q)w^{pq},$$

where w_{pq} is the element of $D'VD$.

Second, for the parametric multinomial distribution, analogue of Cook's distance

$$C_s = (\hat{\theta}-\hat{\theta}(s))'D'VD(\hat{\theta}-\hat{\theta}(s))/P. \qquad (1)$$

$\hat{\theta}(s)$ is the vector of parameter estimates obtained from the conditional distribution given that no observations fall in cell s – and, more specifically, it is the solution to $\sum_{j\neq s} n_j \partial \log \eta_j^*/\partial\theta_p = 0$ $(p=1,\cdots,P)$ with $\eta_j^* = \eta_j/(1-\eta_s)$, which is the conditional probability of an observation being in cell j , given that the observation is not in cell s . D and V are evaluated at ML estimate $\hat{\theta}$. Using one-step approximation to $\hat{\theta}$ by Pregibon(1981), the equation (1) yields

$$\tilde{C}_s \simeq (r_s^2\hat{h}_s/(1-\hat{\eta}_s-\hat{h}_s))/P. \qquad (2)$$

The values close to one of the leverage defined as

$$L_s = \hat{h}_s/(1-\hat{\eta}_s). \qquad (3)$$

worsen the approximation (2).

3 Numerical Examination

First, let us explain the notation appeared in the following tables. $M(MLe)$, $V(MLe)$, and $M(\ a\ .\ v\ .)$ represent, respectively, the mean of $MLe\acute{s}$, the variance of $MLE\acute{s}$and the mean of the asymptotic variance of $MLE\acute{s}$ based on all simulations' evaluations.

1. Example 1

Table 1 and Table 2 represent the results of the simulation of two artificial models with n =4 and m =2 . In this case we are interested in the comparison of the sample variances of estimates based on simulation with thier asymptotic variances at ML estimates. The difference between model 1 and model 2 is the conditional probabilities of the fourth item (n=4) given each latent class(i , e ., in model 1,

π_{41}=0.5 and π_{42}=0.4 ; in model 2, π_{41}=0.8 and π_{42}=0.2). The results are summarized as follows:

1. The latent class probability w_t in model 1 seems to be biased, but that in model 2 seems to be unbiased.
2. The estimates of the latent probability w_1 and the conditional probabilities $\{\pi_{it}\}$ in model 2 as a whole are much better than those of model 1.
3. The sample variances of estimates in model 1 are larger than those of model 2 except the variances of π_{41} and π_{42}.
4. The most important thing is, however, that in both model, asymptotic variances at ML estimates give far well approximations to the sample variances of the latent parameters, hence this suggests that the reliability of parameters can be evaluated by their asymptotic variances at ML estimates.

Table 3 shows the approximate values of Cook's distance given by approximation (2), $\tilde{C}$, and the leverages (3), L , for 2^4 response patterns. The values of the leverage indicate that one-step estimates to $\hat{\theta}(i)$ give good approximation to (1), C . In this case the approximate Cook's distances show that the influence by deleting some cell is relatively small in both models.

2. Example 2

Table 4 is the results of universalistic and particularistic values data (see, for example, Hagenaars and McCutcheon, 2002; McCutcheon, 1987), which consists of the 16-celled crosstabulation of the four survey items, examined earlier by Toby and Stouffer (1951), Lazarsfeld and Henry (1968), and Goodman (1974a,1975, 1979). Table 5 shows the approximate Cook's distances and the exact Cook's distances when a cell is deleted, and the standardized residuals and the leverages. The one-step approximates and exact estimates to $\hat{\theta}(i)$ are shown in Table 6. Lazarsfeld and Henry (1968), and Goodman (1974a,1975, 1979). Table 5 shows the approximate Cook's distances and the exact Cook's distances when a cell is deleted, and the standardized residuals and the leverages. The one-step approximates and exact estimates to $\hat{\theta}(i)$ are shown in Table 6.

Tabele 1 Results of Simulation: Sample size = 1000

Number of Simulation = 390

	w_1	π_{11}	π_{21}	π_{31}	π_{41}	π_{12}	π_{22}	π_{32}	π_{42}
Model 1	.6	.6	.7	.7	.5	.4	.3	.2	.4
$M(MLe)$	.564	.613	.731	.717	.508	.397	.290	.219	.397
$V(MLe)$	.0194	.0016	.0055	.0073	.0012	.0031	.0092	.0086	.0020
M($a.v.$)	.0271	.0020	.0084	.0112	.0012	.0029	.0137	.0153	.0018

Table 2 Results of Simulation: Sample size = 1000

Number of Simulation = 496

	w_1	π_{11}	π_{21}	π_{31}	π_{41}	π_{12}	π_{22}	π_{32}	π_{42}
Model 2	.6	.6	.7	.7	.8	.4	.3	.2	.2
$M(MLe)$	.604	.598	.699	.700	.800	.401	.295	.197	.193
$V(MLe)$	.0034	.0006	.0011	.0012	.0015	.0013	.0016	.0021	.0031
M($a.v.$)	.0031	.0006	.0009	.0012	.0014	.0011	.0016	.0021	.0032

Table 3 $\tilde{C}$ and L of Simulation of Model 1 and Model 2

Cell No.	Response	$\tilde{C}$		L	
		Model 1	Model2	Model 1	Model 2
1	1 1 1 1	.523	.580	.787	.830
2	1 1 1 0	.263	.304	.663	.705
3	1 1 0 1	.216	.237	.660	.678
4	1 1 0 0	.075	.083	.430	.430
5	1 0 1 1	.218	.181	.639	.638
6	1 0 1 0	.094	.056	.428	.345
7	1 0 0 1	.125	.055	.501	.340
8	1 0 0 0	.274	.352	.681	.763
9	0 1 1 1	.167	.211	.589	.646
10	0 1 1 0	.157	.067	.566	.338
11	0 1 0 1	.066	.071	.370	.404
12	0 1 0 0	.381	.276	.722	.713
13	0 0 1 1	.056	.081	.345	.446
14	0 0 1 0	.330	.315	.736	.728
15	0 0 0 1	.271	.414	.667	.772
16	0 0 0 0	.805	.866	.846	.884

We shall now investigate the following five points: 1) one-step approximation and exact estimates to $\hat{\theta}(i)$, 2) approximate Cook's distance and exact Cook's distant, 3) the influence on ML estimates by deleting a cell, 4) the reliability on the ML estimates and 5) the standardized residuals.

From the values of leverages of Table 5, the approximate Cook's distances by one-step approximation give good approximations to the exact Cook's distances except cell No. 8, 12 and 14 - 16, for which we could not obtain exact estimates to $\hat{\theta}(i)$ because of the occurrence of improper solutions, which lie outside interval [0,1], in the iterative process of the Fisher scoring method and for which values of leverage in turn are 0.985, 0.991, 0.993, 0.998, and 0.999. Judging from approximate Cook's distance, in particular, it is considered that deletion of cell No. 12 (approximate Cook's distance = 25.24), No. 15 (= 76.46), No. 16 (= 557.6) give considerable effect. On the other hand, since in this data all standardized residuals | r |'s are below 1.5, we can not find out any influential cell from the viewpoint of these residuals. Table 6 represents the one-step and exact estimates to $\hat{\theta}(i)$. Excluding $\hat{\theta}$ (8) , $\hat{\theta}$ (11) , $\hat{\theta}$ (12) , $\hat{\theta}$ (14-16), the one-step approximations as a whole are fairly good for exact estimates $\hat{\theta}(i)$ and thier estimates considerably coincide with ML estimates $\hat{\theta}$ in Table 4 obtained by using full-data, while one-step approximations of $\hat{\theta}$ (8) , $\hat{\theta}$ (11) , $\hat{\theta}$ (12) , $\hat{\theta}$ (14-16) differ from $\hat{\theta}$ significantly. In particular, No. 8 (π_{12}=-0.54), No. 15 (π_{42}=2.31), No. 16 (w_1=2.216, π_{42}=2.62) are possessed of the improper solutions. Hence, from the viewpoint of Cook's distance, we will conclude that cell No.12, 15, 16 among No. 8, 11, 12, 14, 15 and 16 are considered to be influential and that it appears from the asymptotic variances shown in Table 4 that the reliability of π_{42} (asymptotic variance = 0.0086), π_{32} (= 0.0042), π_{22}

(= 0.0040), and w_1 (= 0.0032) are lower than that of the others.

Table 4 Results of Universalistic and Particularistic Values Data

	w_1	π_{11}	π_{21}	π_{31}	π_{41}	π_{12}	π_{22}	π_{32}	π_{42}
MLe	.721	.286	.646	.670	.868	.007	.074	.060	.231
a.v.	.0032	.0016	.0024	.0024	.0015	.0007	.0040	.0042	.0086

Table 5 r, $\tilde{C}$, C, L of Universalistic and Particularistic Values Data

Cell No.	Response	Frequency	r	$\tilde{C}$	C	L
1	1 1 1 1	20	1.329	0.313	0.185	0.614
2	1 1 1 0	2	-0.409	0.007	0.006	0.264
3	1 1 0 1	6	-0.995	0.061	0.058	0.358
4	1 1 0 0	1	-0.268	0.001	0.001	0.153
5	1 0 1 1	9	-0.079	0.0004	0.0004	0.370
6	1 0 1 0	2	0.538	0.007	0.007	0.170
7	1 0 0 1	4	-0.332	0.005	0.005	0.283
8	1 0 0 0	1	0.304	0.683	-	0.985
9	0 1 1 1	38	-1.156	0.315	0.323	0.679
10	0 1 1 0	7	0.255	0.009	0.009	0.558
11	0 1 0 1	25	1.352	0.376	0.391	0.649
12	0 1 0 0	6	-1.382	25.24	-	0.991
13	0 0 1 1	24	0.134	0.004	0.004	0.671
14	0 0 1 0	6	-0.295	1.301	-	0.993
15	0 0 0 1	23	-1.293	76.46	-	0.998
16	0 0 0 0	42	1.418	557.6	-	0.9996

Table 6 Universalistic and Particularistic Values Data

Latent Prob.		w_1	π_{11}	π_{21}	π_{31}	π_{41}	π_{12}	π_{22}	π_{32}	π_{42}
One-Step	$\hat{\theta}(1)$	.746	.231	.604	.629	.845	.003	.051	.035	.183
Exact		.753	.233	.603	.627	.844	.003	.046	.032	.171
One-Step	$\hat{\theta}(2)$	.725	.289	.647	.671	.860	.005	.069	.055	.233
Exact		.725	.289	.647	.671	.860	.005	.070	.055	.233
One-Step	$\hat{\theta}(3)$	.731	.302	.653	.647	.871	.004	.066	.067	.219
Exact		.730	.302	.653	.649	.871	.004	.066	.067	.220
One-Step	$\hat{\theta}(4)$	.724	.287	.646	.668	.864	.006	.071	.058	.229
Exact		.724	.287	.646	.667	.864	.006	.070	.058	.229
One-Step	$\hat{\theta}(5)$	.722	.288	.644	.671	.868	.007	.074	.060	.230
Exact		.721	.288	.644	.671	.868	.007	.074	.060	.230
One-Step	$\hat{\theta}(6)$	.714	.285	.652	.671	.875	.008	.077	.067	.233
Exact		.714	.285	.652	.671	.876	.008	.078	.067	.234
One-Step	$\hat{\theta}(7)$	.726	.289	.640	.664	.868	.007	.072	.059	.223
Exact		.727	.289	.640	.663	.867	.007	.071	.059	.222
One-Step	$\hat{\theta}(8)$	.741	.290	.639	.663	.864	-.054	.069	.056	.223
Exact		-	.-	-	-	-	-	-	-	-
One-Step	$\hat{\theta}(9)$	.729	.260	.688	.710	.886	.015	.074	.062	.242
Exact		.730	.260	.689	.710	.886	.015	.074	.064	.243
One-Step	$\hat{\theta}(10)$	.714	.290	.646	.670	.878	.008	.079	.067	.231
Exact		.714	.290	.645	.670	.878	.008	.079	.067	.231
One-Step	$\hat{\theta}(11)$	.675	.317	.639	.745	.865	.012	.091	.061	.285
Exact		.676	.318	.639	.746	.865	.012	.088	.062	.276
One-Step	$\hat{\theta}(12)$	.474	.327	.636	.748	.892	.015	.935	.079	.289
Exact		-	-	-	-	-	-	-	-	-
One-Step	$\hat{\theta}(13)$	.716	.290	.654	.670	.867	.007	.074	.062	.236
Exact		.716	.290	.654	.670	.867	.007	.074	.063	.236
One-Step	$\hat{\theta}(14)$	.665	.296	.662	.668	.874	.009	.078	.259	.243
Exact		-	-	-	-	-	-	-	-	-
One-Step	$\hat{\theta}(15)$	.118	.338	.698	.743	.862	.014	.093	.078	2.31
Exact		-	-	-	-	-	-	-	-	-
One-Step	$\hat{\theta}(16)$	2.22	.354	.705	.754	.893	.106	.984	.825	2.62
Exact		-	-	-	-	-	-	-	-	-

References

Andersen, E. B. (1992). Diagnostics in Categorical Data Analysis. *Journalof the Royal Statistical Society, Series B* 54, 781-791.

Cook, R. D. and Weisberg, S. (1982). Residuals and Influence in Regression. New York: Chapman and Hall.

Critchley, F. (1985). Influential in principal component analysis. *Biometrika* 72, 627-636.

Green, B. F. (1951). A general solution for the latent class model of latent structure analysis. *Psychometrika* 16, 151-166.

Gibson, W. A. (1955). An extension of Anderson's solution for the structure equations. *Psychometrika* 20, 69-73.

Goodman, L. A. (1974a). Exploratory latent structure analysis using both identifiable and unidentifiable model. *Biometrika* 61, 215-231.

Goodman, L. A. (1975). A new model for scaling response patterns: An application of the quasi-independence concept. *Journal of the American Statistical Association* 70, 755-768.

Goodman, L. A. (1979). On the estimation of parameters in latent structure analysis. *Psychometrika* 44, 123-128.

Hagenaars, J. A. and McCutcheon, A. L. (2002). Applied Latent Class Analysis. Cambridge: Cambridge University Press.

Lazersfeld, P. F. and Henry, N. W. (1968). Latent Structure Analysis. Boston: Houghton Mifflin.

McHugh, R. B. (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika* 21, 83-89.

McCutcheon, A. L. (1987). Latent Class Analysis. Newbury Park, CA: Sage.

Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics* 9, 705-724.

Stouffer, S. A. and Toby, J. (1951). Role conflict and personality. *American Journal of Sociology* 56, 395-406.

Tanaka, Y. and Odaka, Y. (1989). Influential observations in principal factor analysis. *Psychometrika* 54, 475-485.

（2004 年 12 月 1 日　受理）