A MONTE CARLO ASSESSMENT OF THE STABILITY OF LOG-LINEAR ESTIMATES IN SMALL SAMPLES

Mark Evers, Duke University
N. Krishnan Namboodiri, University of North Carolina at Chapel Hill

Any reasonably complex contingency table will frequently contain empty or zero cells, merely due to sampling fluctuations. Of course, the number of zero cells is negatively related to sample size, and positively related to the number of cells in the contingency table. Thus, theoretically, zero cells can be "removed" either by obtaining a larger sample or by collapsing categories of some of the variables. However, the typical situation is one in which the investigator has only one sample of a given size, and in which collapsing the table is an unattractive alternative. Thus, we have the need for techniques to handle contingency tables with empty cells.

In a situation where there are only a few zero cells, Grizzle, Starmer, and Koch (1969) recommend inserting in each empty cell the value 1/r, where r is the number of response categories. For the iterative maximum likelihood procedure developed by Goodman and others, the following procedure has been suggested in the literature. For each model of interest, examine the marginals to be fitted, and discard all models that require fitting one or more zero marginals. This obviously is not a satisfactory strategy since investigators may wish to estimate parameters for models chosen on a priori grounds. Several options are open if the chosen model requires fitting empty marginals. (1) Use the technique advocated by Grizzle, Starmer and Koch, namely add the quantity 1/r to zero cells and analyze the data with their method. (2) Replace zero cells with small numbers, such as 1/r, and analyze the data using the iterative maximum likelihood technique. (3) Follow the strategy suggested in Bishop et al. (1975), chapter 12, which requires the assumption of a priori cell probabilities.

In this paper we examine strategy (2) in an effort to shed light on the resulting biases in parameter estimates. We refer to the small values added to observed zero cells as correction factors. We address the following questions, using the iterative maximum likelihood procedures as programmed in ECTA (Fay and Goodman 1973). First, does the size of the correction factor systematically affect the parameter estimates one obtains? Second, does the number of zero cells in the contingency table, which is closely related to sample size, influence the behavior of these estimates?

STUDY DESIGN

From the 1-in-100 Public Use Sample (PUS) of the 1970 U.S. Census, we first obtained data for about 219,000 women aged 14 to 44 years. From this data set, we created a four-way contingency table of children ever born by education by race by age. In this table, children ever born had four categories (0, 1, 2-4, 5+), education had three categories (less than 12 years, 12 years, more than 12 years), age had three categories (14-24, 25-34, 35-44), and race had two categories (white, nonwhite), thus giving a table with 72 cells. We specified a hierarchical model, which can be described in terms of the following three-way marginals to be fitted: children ever born by education by race, children ever born by age by race, and education by age by race. This model has 24 degrees of freedom and has a total of 48 independent parameters. In this paper, we examine only the 14 parameters which had the largest estimated values. Table 1 shows these parameter estimates, which we term the full sample estimates, since they are based on the full sample of women from the PUS.

From this full sample of women, we drew several sets of independent random samples: 100 samples of size 250, 100 samples of size 500, and 100 samples of size 1000. For each of these 300 subsamples, we constructed a contingency table with dimensions and categories identical to the table for the full sample of women described above. Every one of these contingency tables contained a number of empty cells, ranging from a minimum of 6 to a maximum of 37. For each of the subsample contingency tables, we used three different correction factors to replace the zero cells--.02, .2, and .5--and we used ECTA to obtain parameter estimates for the model that was fitted to the full sample of data. Thus, this design systematically varies sample size and correction factors, although the three different correction factors were applied to the same set of data. Because there are 100 samples in each set, we also have a reasonable amount of variation in the number of zero cells in each set.

The particular model we chose to fit to these sets of data did not fit well for the full sample of women. The chi-square value was 3304, which with 24 degrees of freedom has a proability of less than .001. It is therefore likely that whatever variation in the parameter estimates that we observe for the different subsamples may in fact be partly attributable to some unknown quantity of specification error. In order to deal with this problem, we took the expected cell counts based on the model fitted to the full sample, and simulated data which paralleled the same study design that was used for the data from the PUS. That is, we simulated 100 samples of each of the three sample sizes, and for each sample, applied each of the three correction factors, and used ECTA to obtain parameter estimates for our model.

RESULTS

Table 2 shows how the estimates of R, the main effect due to race, vary across sample size, correction factor, and number of zero cells in

the contingency table. Results are shown separately for the simulated data, and for the data drawn from the PUS. The bias of the estimates is calculated as the mean value of subsample estimates minus the full sample value. Thus, the value of .508 in the table (top of column 5) refers to the bias for the 61 samples of size 250 in the simulated set, which have between 19 and 28 zero cells, and which have the correction factor of .02 added to the zero cells. For this set of data, the bias is .508, indicating that the mean of the small sample parameter estimates was .508 higher than 1.108, the full sample value for R. The standard deviation for this group of 61 estimates is .273.

There are several patterns for both the bias and the standard deviation which deserve to be noted, since these are similar to the patterns for the other estimates we examined.

First, the correction factor is related to the bias in the following way: overall, the .02 correction tends to produce a positive bias, the .5 correction tends to produce a negative bias, and the .2 correction tends to produce the smallest bias, which hovers close to zero.

This finding makes sense, since, other things being equal, a large increment to zero cells would reduce the heterogeneity of a table and attenuate the value of an effect or a relationship. Hence, a large increment such as .5 would underestimate a positive effect and give a negative bias. On the other hand, a very small correction factor such as .02 clearly overestimates the effect, giving a positive bias.

The second observation about the pattern of bias in Table 2 is that for the .02 and .5 corrections, the amount of bias becomes smaller with increasing sample size. This apparent effect of sample size is most likely due to the number of zero cells in the table, which is strongly and negatively related to sample size. Other things being equal, an increase in the number of zero cells must be offset by larger entries in the remaining cells, giving a larger value for an effect or relationship. Since larger samples have fewer zero cells, we would expect the estimates to tend toward the full sample estimates. This effect is clear for the .02 correction factor, since the bias, or difference between the subsample estimates and the full sample estimate, becomes smaller with increasing sample size.

However, for the .5 correction factor, where the mean of the subsample estimates is consistently less than the full sample value, the bias becomes less negative with increasing sample size. Thus, the subsample estimates are getting larger with sample size, rather than smaller as we would predict by knowing the number of zero cells alone. We argue that the observed trend is due to the attenuating effect of the .5 correction factor. For larger samples, where there are fewer zero cells, there is less chance for this increment to attenuate the size of the estimates.

The third observation about the pattern of bias in Table 2 concerns the effect of zero cells, which we can detect by looking at the trend of bias within sample size. The bias generally becomes more positive as the number of zero cells increases, which gives support to the earlier argument that an increase in the number of zero cells will tend to increase the size of the estimate. Moreover, this effect of the number of zero cells is a good deal stronger for the .02 increment than for the .2 increment, and weakest of all for the .5 increment. It seems that the effect of the number of zero cells on the estimates simply cannot operate as strongly when the increment to these zero cells is larger, but the effect is very clear when the increment is close to zero.

The fourth observation about the pattern in Table 2 concerns the standard deviations of the estimates. We find a strong and negative effect of the size of the correction factor on the magnitude of the standard deviation. This finding is expected, since we know that larger corrections give the estimates more stability.

The last observation about the table is that we can find no major or systematic differences between the simulated data and the data from the Public Use Sample--specification error has no apparent effect on the patterns we observe.

The relation of the size of the estimates, sample size, and the correction factor, is shown in more detail in Table 3, for estimates of R based on the Public Use Sample. For this effect, the correction factor of .2 is likely to give the least bias for the two smaller sample sizes. Indeed, for sample size 250, the .02 and .5 increments do not approximate the full sample estimate of 1.108 for any of the 100 samples. Moreover, the .02 increment yields what most investigators would consider an unacceptably large amount of variation in the sampling distribution of the estimate. If one is willing to tolerate the slightly large standard deviation for the .2 correction factor, this correction yields estimates with relatively small bias, regardless of sample size. That the estimates based on a sample size of 250 can be this good is quite surprising, since the number of zero cells is so large, ranging from 19 to 37 out of a total of 72 cells in the table, and since the average frequency per cell is only 3.5.

Thus far we have considered only one parameter estimate out of the 14 we are examining here. One would naturally ask whether the findings just described can be generalized to the other effects, particularly where they are smaller in value than the estimate of R we have just discussed. The answer is that, generally, we find the same pattern for other effects. Evidence in support of this answer is found in Table 4, which shows the pattern of bias for four other parameters estimates, which differ markedly in size from one another. Close inspection of the table will show that the relationship of bias to sample

size follows the same pattern as we described earlier for the estimate of R. Regarding the correction factor, the value of .2 generally gives the least bias. In contrast to the finding for the estimate of R, this pattern holds even at the largest sample size.

In order to more systematically assess the apparent amount of bias that is linked with the three correction factors, we examined the relative amount of bias for each of the 14 parameter estimates we are considering. For each sample size and for both simulated and real data, we compared the amount of bias that resulted from using each of the three correction factors, and ranked the three factors as yielding high, medium, or low bias, for each estimate. The results of this tally are shown in Table 5. Across all sample sizes, the .2 correction factor consistently is the least likely of the three correction factors to give the highest amount of bias, and in all except the simulated data of sample size 250, the .2 correction factor is most likely to give the least bias. The .02 and .5 correction factors are both very likely to give estimates with a high degree of bias.

The results reported here, of course, concern only one method of dealing with zero cells in contingency tables. We are currently undertaking a Monte Carlo investigation to compare the bias of the estimates and validity of goodness-of-fit tests associated with the correction procedure described in this paper with those associated with the "pseudo-Bayes" procedures described in chapter 12 of Bishop et al. (1975).

REFERENCES

[1] Bishop, Yvonne M. M., Stephen E. Fienberg, and Paul W. Holland. 1975. Discrete Multivariate Analysis: Theory and Practice. Cambridge, Mass.: MIT Press.

[2] Fay, R. and Leo A. Goodman. 1973. "Everyman's Contingency Table Analyzer (ECTA)." Department of Sociology, University of Chicago.

[3] Grizzle, James E., C. Frank Starmer, and Gary G. Koch. 1969. "Analysis of Categorical Data by Linear Models." Biometrics 25 (September): 489-504.

ACKNOWLEDGMENTS

We wish to thank Sharon Poss for extensive programming support, Karen Bothel, Richard Caston, Miltiades Damanakis, Dennis Gilligan, Eugenia Hatley, and Cathie Mayes Hudson for research assistance, and Valerie Hawkins for typing the manuscript. The research was supported by grant SOC76-02100 from the National Science Foundation.

TABLE 1

SELECTED PARAMETER ESTIMATES FOR MODEL FITTED TO CONTINGENCY TABLE BASED ON 1-in-100 PUBLIC USE SAMPLE

Description of Parameter		Full Sample Value (λ)
C_3:	C, third component	1.0214
A_1:	A, first component	-.5808
A_2:	A, second component	.3472
R	R	1.1083
E_2:	E, second component	.3613
E_1R:	E x R, first component	-.3136
C_1A_1:	C x A, first component	1.0385
C_2A_1:	C x A, second component	.9047
C_1A_2:	C x A, fourth component	-.3787
C_2A_2:	C x A, fifth component	-.2622
C_1E_1:	C x E, first component	-.4991
C_2E_1:	C x E, second component	-.2654
C_1A_1R:	C x A x R, first component	.3232
C_3A_1R:	C x A x R, third component	-.2105

Note: C = children ever born, A = age, E = education, R = race.

TABLE 2

BIAS AND STANDARD DEVIATION OF ESTIMATES OF R[a], BY CORRECTION FACTOR, SAMPLE SIZE, TYPE OF DATA, AND NUMBER OF ZERO CELLS

Sample Size	Type of Data	No. of 0 Cells	Correction Factor .02 Bias[b]	.02 Std. Dev.	.20 Bias	.20 Std. Dev.	.50 Bias	.50 Std. Dev.	(N)
250	Simulated	19-28	.508	.273	-.117	.086	-.325	.056	(61)
		29-31	.876	.179	.005	.060	-.261	.043	(29)
		32-37	.985	.167	.018	.080	-.260	.044	(10)
250	PUS	19-28	.596	.290	-.108	.056	-.316	.043	(14)
		29-31	.733	.192	-.051	.058	-.293	.040	(36)
		32-37	.980	.197	.021	.067	-.260	.041	(50)
500	Simulated	13-19	.248	.207	-.019	.081	-.145	.055	(50)
		20-21	.448	.199	.056	.070	-.108	.050	(24)
		22-27	.546	.185	.067	.058	-.117	.037	(26)
500	PUS	13-19	.268	.173	-.017	.065	-.150	.046	(26)
		20-21	.460	.231	.059	.088	-.103	.063	(30)
		22-27	.672	.236	.119	.060	-.079	.040	(44)
1000	Simulated	6-11	.055	.131	-.014	.071	-.061	.056	(47)
		12-13	.178	.114	.050	.056	-.025	.047	(29)
		14-19	.335	.172	.114	.081	.006	.056	(24)
1000	PUS	6-11	.074	.145	.002	.068	.050	.051	(29)
		12-13	.185	.129	.049	.068	-.022	.054	(34)
		14-19	.251	.145	.088	.060	-.003	.046	(37)

[a]Full sample value for R: $\lambda = 1.108$
[b]Bias = Mean value of subsample estimates minus full sample value.

TABLE 3

FREQUENCY DISTRIBUTION OF SUBSAMPLE ESTIMATES OF R[a], BY CORRECTION FACTOR AND SAMPLE SIZE, PUS DATA

Size of Estimate	Sample Size 250 .02	250 .20	250 .50	500 .02	500 .20	500 .50	1000 .02	1000 .20	1000 .50
0.60-0.79	0	0	29	0	0	0	0	0	0
0.80-0.99	0	19	71	0	2	45	5	3	8
1.00-1.19	0	76	0	5	59	55	30	65	91
1.20-1.39	2	5	0	19	39	0	39	32	1
1.40-1.59	8	0	0	31	0	0	25	0	0
1.60-1.79	17	0	0	20	0	0	1	0	0
1.80-1.99	24	0	0	15	0	0	0	0	0
2.00 +	49	0	0	10	0	0	0	0	0
Total	100	100	100	100	100	100	100	100	100
Mean	1.945	1.085	0.828	1.611	1.173	1.003	1.278	1.158	1.085
Bias[b]	.837	-.023	-.280	.503	.065	-.105	-.170	.050	-.023
Std. Dev.	.252	.078	.046	.273	.089	.057	.154	.073	.053

[a]Full sample value for R: $\lambda = 1.108$
[b]Bias = Mean value of subsample estimates minus full sample value.

TABLE 4

BIAS[a] FOR FOUR PARAMETER ESTIMATES OF DIFFERENT SIZE, BY SAMPLE SIZE, TYPE OF DATA, AND CORRECTION FACTOR

Sample Size	Type of Data	Correction Factor	Parameter C_3 (λ=1.021)	C_2A_1 (λ=.905)	C_1E_1 (λ=-.499)	E_1R (λ=-.314)
250	Simulated	.02	.244	.100	-.178	-.136
		.20	-.207	-.319	.044	.085
		.50	-.365	-.473	.138	.152
250	PUS	.02	.115	.155	.113	-.122
		.20	-.234	-.303	.097	.121
		.50	-.354	-.422	.119	.189
500	Simulated	.02	.254	.170	-.107	-.121
		.20	-.096	-.178	.010	.044
		.50	-.241	-.328	.073	.107
500	PUS	.02	.227	.323	.035	-.096
		.20	-.120	-.155	.046	.066
		.50	.253	-.340	.057	.129
1000	Simulated	.02	.161	.156	-.065	-.057
		.20	-.031	-.075	-.011	.017
		.50	-.129	-.193	.027	.057
1000	PUS	.02	.197	.164	-.011	.122
		.20	-.013	-.085	.028	.085
		.50	-.116	-.214	.053	.072

[a]Bias = Mean value of subsample estimates minus full sample value.

TABLE 5

FREQUENCY WITH WHICH DIFFERENT CORRECTION FACTORS RESULT IN HIGH, MEDIUM, OR LOW BIAS ACROSS 14 PARAMETER ESTIMATES

Type of Data	Correction Factor	Sample Size 250 Amount of Bias High	Medium	Low	500 Amount of Bias High	Medium	Low	1000 Amount of Bias High	Medium	Low
Simulated	.02	3	2	9	7	1	5	5	6	3
	.20	0	9	5	0	5	9	1	2	11
	.50	11	3	0	7	7	0	9[a]	4	1[b]
PUS	.02	4	4	6	5	5	4	8	3	3
	.20	0	6	8	0	6	8	0	6	8
	.50	10	4	0	9	4	1	6	5	3

[a]For one parameter, correction factors of .02 and .50 tied for "high" bias, .20 was assigned "low" bias.

[b]For one parameter, correction factors of .20 and .50 tied for "low" bias, .02 was assigned "high" bias.