DOCUMENT RESUME

ED 336 047 HE 024 869

AUTHOR	Saupe, Joe L.
TITLE	A Spreadsheet for a 2 x 3 x 2 Log-Linear Analysis. AIR 1991 Annual Forum Paper.
PUB DATE	May 91
NOTE	21p.; Paper presented at the Annual Forum of the Association for Institutional Research (31st, San Francisco, CA, May 26-29, 1991).
PUB TYPE	Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Computer Oriented Programs; *Computer Software; Higher Education; *Institutional Research; Microcomputers; *Spreadsheets; *Statistical Analysis
IDENTIFIERS	*AIR Forum; *Log Linear Models; Multidimensional Models; University of Missouri Columbia

ABSTRACT

This paper describes a personal computer spreadsheet set up to carry out hierarchical log-linear analyses, a type of analysis useful for institutional research into multidimensional frequency tables formed from categorical variables such as faculty rank, student class level, gender, or retention status. The spreadsheet provides a concrete vehicle for presenting the general nature, as well as the computational details of a log-linear procedure. The program was prepared in Symphony but can be translated to Lotus 123 or replicated readily with other spreadsheet software. It includes 11 sections or pages: one for entering observed frequencies (the only input required), one for each of 8 applicable log-linear models, one for iteratively-derived estimates of expected values for the model for the third second-order interaction and one for summary tables. The spreadsheet could be used with non-hierarchical models by copying the page for one of the models analyzed onto an empty spreadsheet range, and then changing the formulas which calculate the expected frequencies for the added model. It would also be possible to add a section to each model page of the spreadsheet in which "standardized residuals" were displayed. Five references and five tables complement the text. (JB)

A SPREADSHEET FOR A 2 X 3 X 2 LOG-LINEAR ANALYSIS

Joe L. Saupe
Professor, Higher and Adult Eǰucations and Foundations
University of Missouri-Columbia

301 Hill Hall
University of Missouri-Columbia
Columbia, MO 65211

(314) 882-5123

Paper presented at the Thirty-First Annual Forum
The Association for Institutional Research
The Westin St. Francis, San Francisco, California
May 26-29, 1991.

AIR

for Management Research, Policy Analysis, and Planning

This paper was presented at the Thirty-First Annual Forum of the Association for Institutional Research held at The Westin St. Francis, San Francisco, California, May 26-29, 1991. This paper was reviewed by the AIR Forum Publications Committee and was judged to be of high quality and of interest to others concerned with the research of higher education. It has therefore been selected to be included in the ERIC Collection of Forum Papers.

Jean Endo
Chair and Editor
Forum Publications Editorial
Advisory Committee

Abstract

A SPREADSHEET FOR A 2 X 3 X 2 LOG-LINEAR ANALYSIS

A personal computer spreadsheet set up to carry out hierarchical log-linear analyses, association or logit, is described. The spreadsheet provides a concrete vehicle for introducing the nature and computational details of log-linear analysis and is an alternative to advanced statistical packages in carrying out the required computations. The spreadsheet includes 11 sections or pages, one for entering observed frequencies (the only inputs required), one for each of 8 applicable log-linear models, one for iteratively-derived estimates of expected values for the model for the third second-order interaction and one of summary tables. Approaches to converting the spreadsheet for use with non-hierarchical models and with other designs are discussed.

A Spreadsheet for a 2 x 3 x 2 Log-Linear Analysis

The applicability of log-linear analysis to problems for institutional research has been described by Hinkle and McLaughlin (1984). Hinkle, McLaughlin and Austin (1988) and Moline et al (1989) who have also provided descriptions of the statistics of the log-linear approach to data analysis. Log-linear models are used with multidimensional frequency tables formed from categorical variables. Faculty rank, student class level, student enrollment category (entering freshman - transfer - continuing), gender, ethnic classification, major - non-major, and retention status are illustrations of categorical variables with which institutional researchers deal and which make log-linear analysis a tool which belongs in the IR repertoire.

While the basic notion of log-linear models and the nature of the two types of such models, association and logit, are not complex, the multidimensionality of the analysis can make the approach difficult to comprehend. Also the volume of the "results" for a non-complex log-linear analysis can make the technique forboding.

The purpose of this paper is to describe a personal computer spreadsheet set up to carry out a 2 x 3 x 2 hierarchical log-linear analysis. The spreadsheet is applicable to the analysis of both association and logit models. The spreadsheet is intended to serve two functions. First, it provides a concrete vehicle for presenting the general nature, as well as the computational details, of the log-linear procedure. The first function is pedagogical. Secondly, the spreadsheet carries out the calculations and provides summary tables for the analysis upon the input of the observed frequencies of the 12 cells of the 2 x 3 x 2 table. The second function is a practical one. While the 2 x 3 x 2 design is a specific case, the spreadsheet developed for this case provides a model for developing spreadsheets for other designs.

With regard to its practical use, the spreadsheet approach to log-linear analysis is not being offered as a substitute for the log-linear routines in such statistical packages as SAS, SPSSX, BMDP and SYSTAT. For relatively straightforward designs, such as the 2 x 3 x 2 one, however, once one has the observed frequencies, the spreadsheet provides an alternative to the statistical package.

The spreadsheet was prepared in SYMPHONY, but can be translated to LOTUS 123. Its design is sufficiently straightforward that it can be replicated quite readily with other spreadsheet software.

The spreadsheet was developed to analyze categorical data from a cohort of entering college students. Thus, the table headings used apply to those data. They could, of course, be changed to describe the categories used with some other set of data or could be made generic.

The three categorical variables used in the illustration are (1) type of institution attended for the student's first year in college, categorized as 2-year or 4-year, (2) type of institution attended during the student's second year in college, categorized as (transferred to) two-year, (transferred to) four-year or (remained at) same college or university attended the first year and (3) did student complete a college preparatory curriculum in high school?, categorized as yes or no. While these data rather clearly call for a logit analysis, they are used to illustrate the association, as well as, the logit application of the spreadsheet.

Log-linear analysis is similar to the familiar chi-square analysis of a two-way frequency table, except that (1) normally more than two categorical variables are involved and (2) in addition to testing for contingencies or associations between pairs (or sets) of variables, goodness of fit questions are asked about marginal total frequencies. For an association analysis, there is no dependent variable; all of the variables are considered to be independent, as in factor analysis. To carry out an association analysis, models, from simple to complex, are sequentially fitted to the observed data. The simplest model to fit the data is considered to be the solution and no more complex models are tested.

The logit variety of log-linear analysis involves a dependent variable. The purpose of the analysis is to test for dependencies of this variable upon the other, independent, variables and combinations (interactions) thereof. In the spreadsheet to be described, the two independent variables are called "Explanatory Variables" and the dependent variable is called the "Response Variable."

Page 1 and Observed Frequencies

Figure 1 shows the first "page" of the spreadsheet. Each "page" includes 60 rows, including for page 1 (but not the others) some blank rows at the bottom. Thus, one may "page down" three times to move from page to page. Of course, not all of a page is visible on the screen at any one time (unless one would happen to use software that permits the display of 60 rows). In the study for which the spreadsheet was developed there were several dichotomous dependent or response variables (Honey, 1991). Thus, there is a row at the top of the page for entering the name of the specific response variable being analyzed by the spreadsheet.

The observed frequencies are entered in the twelve (2 x 3 x 2) cells of Part A of the first page. There were 17 students who began college at a two-year institution, enrolled for their second year in college in another two-year institution and had completed the college preparatory curriculum in high school; 51 began at a two-year institution, transferred to a four-year one for their second year and had completed the college preparatory curriculum; and 421 began at a two-year college, stayed at that college for their second year and had completed the college preparatory curriculum. The largest number, 4,618, began at a four-year college or university, remained at that college or university for their second year and had not completed the college preparatory curriculum in high school.

These twelve frequencies are the only data that are entered to the spreadsheet. The values in the remaining three sections of this page and all of the values on the other ten pages of the spreadsheet are derived by formulas from these twelve numbers.

In Part B of the first page the observed frequencies are converted to proportions by formulas which divide the respective values in Part A by the total of the values in Part A. These proportions are not a central part of the

```
                 ILLUSTRATIVE LOG-LINEAR ANALYSIS
             ("V" stands for some dichotomous variable)

VARIABLE:      College Prep Curriculum? -- "1", Yes;  "0", No

                          0. OBSERVED FREQUENCIES

A. OBSERVED FREQUENCIES                     D. MARGINAL TOTALS
                                                      Obs F      Prop

Response ------Explanatory Variables-----   m(...) ...   11,509      1.00
Variable F'st-Year  S'nd-Year Institution
   (V)   Inst. (F)         (S)              m(F..) 2..    1,918      0.17
                    2-Year  4-Year    Same         4..    9,591      0.83
-----------------------------------------
   "1"   2-Year        17      51      421  m(.S.) .2.      723      0.06
         4-Year       240     263     3729         .4.      752      0.07
                                                   .S.   10,034      0.87
   "0"   2-Year        47     116     1266
         4-Year       419     322     4618  m( .V) ..1    4,721      0.41
-----------------------------------------          ..0    6,788      0.59

B. PROPORITONS                              m(FS.) 22.       64      0.01
                                                   24.      167      0.01
Response ------Explanatory Variables-----          2S.    1,687      0.15
Variable F'st-Year  S'nd-Year Institution          42.      659      0.06
   (V)   Inst. (F)         (S)                     44.      585      0.05
                    2-Year  4-Year    Same         4S.    8,347      0.73
-----------------------------------------
   "1"   2-Year      0.00    0.00     0.04  m(F.V) 2.1      489      0.04
         4-Year      0.02    0.02     0.32         4.1    4,232      0.37
                                                   2.0    1,429      0.12
   "0"   2-Year      0.00    0.01     0.11         4.0    5,359      0.47
         4-Year      0.04    0.03     0.40
-----------------------------------------   m(.SV) .21      257      0.02
                                      1.00         .41      314      0.03
C. PROPORTIONS OF "1"s FOR V                       .S1    4,150      0.36
                                                   .20      466      0.04
------Explanatory Variables-----                   .40      438      0.04
          S'nd-Year Institution                    .S0    5,884      0.51
F'st-Year          (S)
Inst. (F)  2-Year  4-Year     Same  Totals
-----------------------------------------
2-Year       0.27    0.31     0.25    0.25
4-Year       0.36    0.45     0.45    0.44

Total        0.36    0.42     0.41    0.41
-----------------------------------------
```

FIG. 1. Page 1 of spreadsheet -- observed frequencies.

log-linear analysis, but may be useful in attaching meaning to the frequencies. In the lower right hand corner of Part B appears a 1.00 which comes from a formula summing the twelve proportions. The sum simply serves as a check on the formulas within the Part B table.

Part C of the first page is included for the logit analysis of the data. It provides proportions of observations classified by the two independent or explanatory variables which had "1s" on the dependent or response variable. Marginal totals are included. In the logit analysis the proportions in this table are the ones being compared in the fashion of a two-way analysis of variance. Each formula in a cell of Part C is a number in Part A divided by that number plus another one. For example, in the first cell, .27 = (17)/(17 + 47).

The marginal totals of observed values are given in Part D of the first page. The row headings of Part D identify the specific totals which are displayed, using dot notation. The grand total is identified as "m(...) ...." The number of observation for which the value of variable F is 2 is identified by "m(F..), 2.."; with the illustrative data this is the total number of students who began at two-year colleges. The formulas in the cells of the Obs F column are sums of values in the table in Part A.

The marginal totals of the observed frequencies are converted to proportions and are shown in the column headed "Prop." A formula in this column is the value in the cell to the left divided by the value in the cell labeled m(...). As with the Part B table these proportions are not critical to the log-linear analysis.

The Log-Linear Model and Its Estimation

The spreadsheet is designed to carry out a hierarchical, association log-linear analysis. What this means is that successive models generate expected frequencies for the twelve cells. The observed frequencies are compared with the expected frequencies produced by each model in order to determine whether or not the model provides a satisfactory explanation for the observed frequencies. The models are cumulative and the procedure stops with the first model which is found to satisfactorily explain the data. The complete log-linear model can be expressed as,

$$f_{fsv} = e^{m}, \qquad (1)$$

where $m = u + a_{f..} + a_{.s.} + a_{..v} + b_{fs.} + b_{f.v} + b_{.sv} + c_{fsv}$,

and f_{fsv} = the population value for cell fsv,

u = the population value of the grand mean,

the a's = effects for categories of individual variables,

the b's = effects for combinations of categories of pairs of variables, and

c = the effect for an individual cell, for a combination of all three variables.

The first model that is tested is,

$$m = u$$

This is called the null model, because all of the effects are zero. If the data do not fit this model, the next one tested is,

$$m = u + a_{f..}.$$

Next, $m = u + a_{f..} + a_{.s.}$

is tested and the procedure continues until a model which fits the data is found. If no other model fits the data, the full model which includes all of the effect terms is the solution.

When the model which fits the data is found, it may be desirable to estimate the parameters of the model. Parameter estimates are obtained for the full model as follows (Marascuilo & Busk, 1987). The estimation procedures are the same for the other models, except that values of effects not included in the model are zero. First, because m is the natural logarithm of f_{fsv},

$$\ln f_{fsv} = u_{...} + a_{f..} + a_{.s.} + a_{..v} + b_{fs.} + b_{f.v} + b_{.sv} + c_{fsv}. \qquad (2)$$

So, the natural logarithm of each cell is taken and values such as the following ones are calculated (using S_i as the summation sign, with i indicating the variable over which the sum is taken):

$$Y_{...} = (1/FSV)S_f S_s S_v \ln f_{fsv}, \tag{3}$$

$$Y_{f..} = (1/SV)S_s S_v \ln f_{f..}, \tag{4}$$

$$Y_{fs.} = (1/V)S_v \ln f_{fs.}, \text{ and} \tag{5}$$

$$Y_{fsv} = \ln f_{fsv}. \tag{6}$$

In these formulas F, S, and V are the number of categories of the respective variables. Values of $Y_{.s.}$, $Y_{..v}$, $Y_{f.v}$, and $Y_{.fv}$ are also calculated using corresponding formulas.

Estimates of the parameters are then calculated by formulas such as the following:

$$u = Y, \tag{7}$$

$$a_{f..} = Y_{f..} - Y, \tag{8}$$

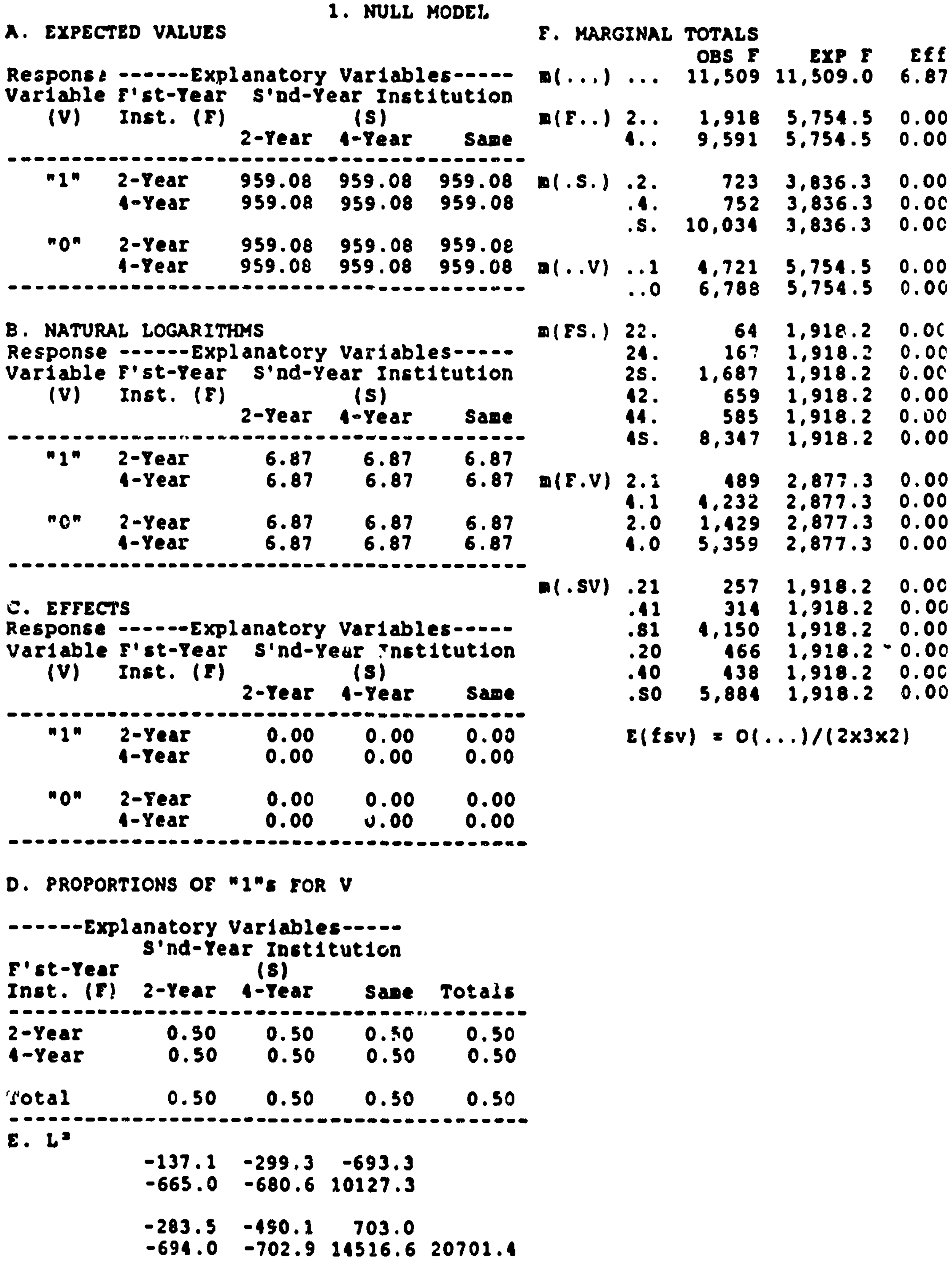
$$b_{fs.} = Y_{fs.} - Y_{f..} - Y_{.s.} + Y, \text{ and} \tag{9}$$

$$c_{fsv} = Y_{fsv} - Y_{fs.} - Y_{f.v} - Y_{.sv} + Y_{f..} + Y_{.s.} + Y_{..v} - Y. \tag{10}$$

As before, the formulas for calculating $a_{.s.}$, $a_{..v}$, $b_{f.v}$ and $b_{.sv}$ are parallel to these.

Page 2 and The Null Model

Figure 2 includes the page devoted to the null model. To repeat, all of the values on this page are the results of formulas in the spread sheet cells. The formulas use values in cells on either page 1 or values, already calculated, on this page.



1. NULL MODEL

A. EXPECTED VALUES

Response	------Explanatory Variables-----			
Variable	F'st-Year	S'nd-Year Institution		
(V)	Inst. (F)		(S)	
		2-Year	4-Year	Same
"1"	2-Year	959.08	959.08	959.08
	4-Year	959.08	959.08	959.08
"0"	2-Year	959.08	959.08	959.08
	4-Year	959.08	959.08	959.08

B. NATURAL LOGARITHMS

Response	------Explanatory Variables-----			
Variable	F'st-Year	S'nd-Year Institution		
(V)	Inst. (F)		(S)	
		2-Year	4-Year	Same
"1"	2-Year	6.87	6.87	6.87
	4-Year	6.87	6.87	6.87
"0"	2-Year	6.87	6.87	6.87
	4-Year	6.87	6.87	6.87

C. EFFECTS

Response	------Explanatory Variables-----			
Variable	F'st-Year	S'nd-Year Institution		
(V)	Inst. (F)		(S)	
		2-Year	4-Year	Same
"1"	2-Year	0.00	0.00	0.00
	4-Year	0.00	0.00	0.00
"0"	2-Year	0.00	0.00	0.00
	4-Year	0.00	0.00	0.00

D. PROPORTIONS OF "1"s FOR V

------Explanatory Variables-----				
	S'nd-Year Institution			
F'st-Year		(S)		
Inst. (F)	2-Year	4-Year	Same	Totals
2-Year	0.50	0.50	0.50	0.50
4-Year	0.50	0.50	0.50	0.50
Total	0.50	0.50	0.50	0.50

E. L^2

-137.1	-299.3	-693.3	
-665.0	-680.6	10127.3	
-283.5	-490.1	703.0	
-694.0	-702.9	14516.6	20701.4

F. MARGINAL TOTALS

		OBS F	EXP F	Eff
m(...)	...	11,509	11,509.0	6.87
m(F..)	2..	1,918	5,754.5	0.00
	4..	9,591	5,754.5	0.00
m(.S.)	.2.	723	3,836.3	0.00
	.4.	752	3,836.3	0.00
	.S.	10,034	3,836.3	0.00
m(..V)	..1	4,721	5,754.5	0.00
	..0	6,788	5,754.5	0.00
m(FS.)	22.	64	1,918.2	0.00
	24.	167	1,918.2	0.00
	2S.	1,687	1,918.2	0.00
	42.	659	1,918.2	0.00
	44.	585	1,918.2	0.00
	4S.	8,347	1,918.2	0.00
m(F.V)	2.1	489	2,877.3	0.00
	4.1	4,232	2,877.3	0.00
	2.0	1,429	2,877.3	0.00
	4.0	5,359	2,877.3	0.00
m(.SV)	.21	257	1,918.2	0.00
	.41	314	1,918.2	0.00
	.S1	4,150	1,918.2	0.00
	.20	466	1,918.2	0.00
	.40	438	1,918.2	0.00
	.S0	5,884	1,918.2	0.00

E(fsv) = O(...)/(2x3x2)

FIG. 2. Page 2 of spreadsheet -- the null model.

The observed frequencies for the marginal totals in Part F are copied from the corresponding section of page 1. The expected values in Part A are calculated from the observed marginal totals. In the case of the null model, all of the effects, except for the overall total, u, are zero. This means that the expected frequencies for all of the cells are the same. Thus the expected values in the Part A table are each calculated as the total observed frequency divided by the number of cells, 11,509/12. This formula appears as a label at the foot of Part F.

The marginal expected frequencies in Part F are calculated by summing cell expected values in Part A, using formulas that are parallel to those on page 1 used to find the marginal totals for the observed frequencies.

Part B is included as a storage place for the natural logarithms of the cell expected values which are used to calculate the effects of the model being tested as outlined in formulas (2) through (10). As indicated in these formulas, the natural logarithms are first converted to "Ys," formulas (3) through (6), which are then converted to the estimates of effects, formulas (7) through (10). As formula (6) shows, the Y-value for an individual cell is the natural logarithm of the expected frequency for that cell. Thus, these Y values are contained in Part B of the page. The marginal Y-values are contained in a column of the spreadsheet which is to the right of the Eff column in Part F and are calculated by formulas such as (3), (4), and (5). (These values are not very interesting and are not shown in Figure 2 in order to maximize the size of the print in the figure.)

Estimates of effect sizes, calculated from formulas like (7) through (10), are displayed in Part C and in the Eff column of Part F of the page. Of course, for the null model all effects are zero. The model includes only the "effect" for the grand total, 6.87.

Part D includes proportions of "1s" on the response variable specified by the null model. These proportions are calculated from the expected frequencies in Part A with formulas that parallel those of Part C on page 1. They are relevant to only the analysis of logit models.

In a log-linear analysis observed frequencies are compared to expected frequencies using a statistic which is called L^2 (or G^2 in some sources).

$$L^2 = 2S_{cells}O\ln(O/E) \qquad (11)$$

where O = Observed frequency,

E = Expected frequency,

ln = Natural logarithm,

and S_{cells} is used to indicated summation over all cells.

The calculated value of L^2 normally is very similar to the value of the chi-square statistic calculated from the same observed and expected frequencies. Both of these statistics have chi-square sampling distributions. The L^2 is preferred for log-linear analysis because it is additive. The hypothesis being tested is that the model explains the observed frequencies, that the data fit the model or that the observed frequencies are only chance departures from the expected frequencies produced by the model. If L^2 is small, for example, is less than the 95th percentile of the relevant chi-square distribution, it is concluded that the model fits the data. If L^2 is large, for example, exceeds the 95th percentile of the relevant chi-square distribution, it is concluded that the model does not fit the data.

The L^2 statistic for each cell calculated by formula (11) and the sum of the individual values are displayed in Part E of the spreadsheet page. The null model L^2 is 20,701.4 which is clearly significant The null model does not fit the data.

Pages 3 Through 9 for Additional Models

Pages 3 through 9 of the spreadsheet are set up identically to page 2. except that each contains data for a different model. The full set of models tested and the page which contains the statistics for each model are included in Table 1. The models are named on the basis of the effects included in the model. This naming makes it clear that the log-linear procedure being used is hierarchical.

TABLE 1. Page Number and Formula for Expected Values for Each Model in Hierarchical Analysis

Page	Model	Formula for E(fsv)
2	Null	O(...)/(2x3x2)
3	F	O(f..)/(2x3)
4	F,S	O(f..) x O(.s.)/(2xO(...))
5	F,S,V	O(f..) x O(.s.) x O(..v)/O(...)^2
6	F,S,V,FS	O(fs.) x O(..v)/O(...)
7	F,S,V,FS,FV	O(fs.) x O(f.v)/O(1..)
8	F,S,V,FS,FV,SV	*
9	F,S,V,FS,FV,SV,FSV	O(fsv)

* An iterative procedure is used.

These seven pages of spreadsheet labels and formulas were created, largely, by use of the spreadsheet "copy" command. The contents of all of the model pages of the spreadsheet are identical, except for the formulas in Parts A which calculate the expected values for the respective models and, of course, the formulas for calculating expected values which appear as labels below the Parts F columns.

As one moves from page to page, model to model, through the spreadsheet, the nature of the models with regard to marginal expected frequencies is shown. Specifically, the marginal frequencies which correspond to the model being tested are the same as the observed frequencies. As additional effects are added to the model additional marginal expected frequencies are the same as the corresponding observed frequencies. For the full model on page 9, all corresponding expected frequencies and observed frequencies are the same. In effect, the formulas for calculating cell expected frequencies fix the marginal frequencies for the effects of the model being tested and distribute expected frequencies to the remaining marginals and to the individual cells as equally as possible.

It is interesting to note that as one moves from page to page, through the spreadsheet, values (other than zero) for model effects are added in the Eff column of Parts F of the spreadsheet pages to reflect the nature of the model being analyzed on each page. Part C contain zeros on each page except page 9 which is the full model page.

Iterative Calculation of Expected Frequencies For Model F,S,V,FS,FV,SV

As shown in Table 1, the expected values for all of the models, except the one which includes all three second-order interaction terms, are functions of marginal observed values. The spreadsheet makes use of an iterative procedure described by Fienberg (1980) to estimate the expected values for the model for which marginal observed values cannot be used. Figure 3 displays the portion of the spreadsheet which is used to carry out these calculations and displays the results for the illustrative data. The Feinberg formulas are in the worksheet cells. It may not be clear that convergence at the sixth iteration

ILLUSTRATIVE LOG-LINEAR ANALYSIS

VARIABLE: College Prep Curriculum? -- "1", Yes; "0", No

ITERATIVE CALCULATION OF EXPECTED VALUES FOR MODEL F,S,V,FS,FV,SV

	ITERATION							
CELL	0	1	2	3	4	5	6	Final
221	1.00	32.00	16.32	13.66	13.0	12.92	12.89	12.89
220	1.00	32.00	47.68	53.43	51.0	51.11	51.19	51.19
241	1.00	83.50	42.58	44.46	44.9	44.58	44.64	44.64
240	1.00	83.50	124.42	120.76	122.1	122.40	122.28	122.28
2S1	1.00	843.50	430.11	433.95	435.0	431.50	431.52	431.52
2S0	1.00	843.50	1256.89	1249.08	1252.02	1255.49	1255.42	1255.42
421	1.00	329.50	290.78	243.34	244.5	244.75	244.11	244.11
420	1.00	329.50	368.22	412.57	414.5	414.16	414.81	414.81
441	1.00	292.50	258.13	269.54	268.7	269.01	269.36	269.36
440	1.00	292.50	326.87	317.24	316.3	316.01	315.72	315.72
4S1	1.00	4173.50	3683.09	3716.05	3714.28	3718.24	3718.48	3718.48
4S0	1.00	4173.50	4663.91	4634.92	4632.72	4628.82	4628.58	4628.58

FIG. 3. Spreadsheet range used for iterative calculations for expected values for model F,S,V,FS,FV,SV.

of the Feinberg calculations is satisfactory, but the final estimates of cell expected values in the illustration do produce marginal expected values which differ from the observed marginal frequencies by no more than 0.1.

A Summary Page for the Hierarchical Log-Linear Analysis

Figure 4 shows a section of the spreadsheet which includes a summary of the hierarchical log-linear analysis. All of the entries on the summary page are labels, except for the variable name in the heading, the values of L^2 and the entries in the columns headed "Signif." The variable name is copied from page 1 (see Figure 1). Residual L^2 values are copied from the pages for the respective models, component L^2 values are obtained by subtraction as explained in the following section and the entries in the Significance column come from a lookup function which makes use of an abbreviated chi-square table located in a range adjacent to the summary table. The totals of L^2 values in the two tables come from a summation function.

Component L^2 and Logit Models

The L^2 statistics which are calculated on the individual model pages of the spreadsheet are called residual L^2s. Note in Figure 4 that for the illustrative data the first model which produced a non-significant L^2 was the model which includes all effects, except the one for the triple interaction. The finding is that this is the model which fits the data. Component L^2s, which appear to the right of the residual values on the spreadsheet summary page, are obtained by subtraction of the residual L^2 on the same row from the Residual L^2 on the preceding row. These L^2s reflect the change in L^2 resulting from the addition of the subject effect to the model and, thus, permit evaluation of the contribution of individual factors. A significant component L^2 indicates that the effects for the last factor added to the model are significant. As can be seen in Figure 4, consistent with the findings from the residual L^2s, each of the individual factors, except for the triple interaction, are significant for the illustrative data.

The Summary Table for the results when the data include two independent (explanatory) and one dependent (response) variable, called the logit analysis,

SUMMARY OF HIERARCHICAL LOG-LINEAR ANALYSIS

VARIABLE: College Prep Curriculum? -- "1", Yes; "0", No

A. ASSOCIATION MODELS

	Residual			Component		
MODEL	L2	df	Signif.	L2	df	Signif.
1. Null	20,701.4	11	***			
2. F	15,117.0	10	***	5,584.4	1	***
3. F,S	686.3	8	***	14,430.6	2	***
4. F,S,V	313.1	7	***	373.3	1	***
5. F,S,V,FS	260.4	5	***	52.7	2	***
6. F,S,V,FS,FV	19.8	4	***	240.7	1	***
7. F,S,V,FS,FV,SV	3.5	2	n.s.	16.2	2	***
8. F,S,V,FS,FV,SV,FSV	0.0	-	--	3.5	2	n.s.
			TOTAL	20,701.4	11	

* .010 < P < .050
** .001 < P < .010
*** .010 < P < .050

B. LOGIT MODELS

	Residual			Component		
MODEL	L2	df	Signif.	L2	df	Signif.
1. NULL	260.4	5	***			
2. F	19.8	4	***	240.7	1	***
3. S	3.5	2	n.s.	16.2	2	***
4. FS	0.0	-	--	3.5	2	n.s.
			TOTAL	260.4	5	

* .010 < P < .050
** .001 < P < .010
*** .010 < P < .050

FIG. 4. Summary page of spreadsheet.

appears at the bottom of the summary page of the spreadsheet, shown in Figure 4. In this use of log-linear analysis, there is no interest in effects of individual variables or of interactions which do not involve the dependent variable. In the illustration the null model for the logit analysis is model F,S,V,FS and the residual and component L^2s for this and subsequent models are the same as shown for the association model. The finding is that, with regard to variable V (whether or not the student had completed the college preparatory curriculum), factor F (type of institution entered for the first year of college) and factor S type of institution enrolled in --transferred to -- the second year) are significant, but the interaction of F and S are not. This finding is interpreted on the basis of the proportions in Part D of Figure 1.

Expanding the Spreadsheet to the Non-Hierarchical Case and Adding Standardized Residuals

In hierarchical log-linear analysis, variables are entered, or added to the model, in a predetermined sequence on the basis of theory or some other rationale. The resulting models that are tested do not exhaust the set of possible models which might explain the observed data. Non-hierarchical log-linear analysis is used to analyze frequency data by examining the full set of possible explanatory models. The spreadsheet described here could be extended to the non-hierarchical case fairly easily by copying the "page" for one of the models now analyzed to an empty spreadsheet range, and then changing the formulas in Part A table of expected frequencies to ones which calculate the expected frequencies for the added model. This could be repeated for all possible models.

Also, it would be straightforward to add a section to each model page of the spreadsheet in which "standardized residuals" are displayed. Standardized residuals are values calculated for cells of the frequency table from observed and expected frequencies, which have an approximately normal. distribution and which are used to test post hoc hypotheses about the contributions of individual cells to significant L^2s. Once this section was added for one page, it could be copied to all other model pages using the spreadsheet copy command.

Summary

A personal computer spreadsheet designed to carry out a 2 x 3 x 2 hierarchical log-linear analysis has been described. The analysis is produced by the entry of the twelve observed frequencies. The spreadsheet should be useful to institutional researchers who want an introduction to the methods of log-linear analysis and to those who are familiar with the procedure and would like to carry it out using a personal computer spreadsheet rather than an advanced statistical package. The present spreadsheet could be used as a basis for developing spreadsheets for other log-linear models. Expansions of the present spreadsheet to non-hierarchical models and to the inclusion of standardized residuals were introduced.

References

Fienberg, S.E. (1980). The analysis of cross-classified categorical data. (2nd ed.) Cambridge, MA: MIT Press.

Hinkle, D. E. & McLaughlin, G. W. (1984). Selection of models in contingency tables. Research in Higher Education, 21, 415-423.

Hinkle, D. E., McLaughlin, G. W. & Austin J. T. (1988). Using log-linear models in higher education research. In B. D. Yancy (ed.), Applying statistics in institutional research, New Directions In Institutional Research, No. 58.

Honey, D. A. S. (1991). An exploratory study of differences among groups of Missouri public higher education transfer students. Unpublished doctoral dissertation, Univesity of Missouri-Columbia, Columbia.

Moline, A. E. (Moderator), Applying statistics in institutional research, (Panel). 29th Annual Forum, The Association for Institutional Research, Baltimore, MD, 1989.