APPLICATION OF LOG-LINEAR MODELLING TO HEALTH SURVEY DATA*

R. C. Hanumara, M. H. Branson, University of Rhode Island

I. INTRODUCTION

This paper presents the results of the analysis of data of measures of use and need of health services for a defined population. The data presented were obtained through a statewide household survey conducted by Rhode Island Health Services Research, Inc. (SEARCH) during the early part of 1972.** The sample was a full probability sample with households randomly selected from within each of Rhode Island's thirty-nine cities and towns. Interviews were obtained for 93 percent of the families falling in the sample. Information on disability and utilization of health services was obtained for 3,086 families consisting of 9,383 individuals. More detailed information describing the survey methodology is available in Thornberry et. al (1973). Number of bed days per person per year (measuring need) and number of physician visits per person per year (use) are used in the present study as dependent variables. The data on independent variables, age, sex, and economic status, is also available from the SEARCH survey.

Regression procedures to study the relationship between a "dependent" variable and "independent" variables, when both the dependent and independent variables are quantitative are extensively discussed in the literature, and are widely used in practice. Even if some of the independent variables are qualitative, it is possible to use regression procedures by introducing dummy variables. Furthermore, there are several methods of selecting the models that fit the data (Draper and Smith, 1966) and recent papers in Technometrics. However, when the dependent variable is a qualitative variable, these regression procedures are not directly applicable. We note here the above comments apply equally well to analysis of variance method as regression model techniques encompass a general situation.

One may still want to consider analysis of variance method on the data after applying logarithmic or square root transformation. Normality, independence, and equal variances are the usual assumption required in analysis of variance method, which are not satisfied for the health data even after transformation. There is a large percentage of people with zero counts on both the dependent variables and these present additional difficulties in transformation. One other important concern is the quality of the data itself. A disproportionately large number of people reported to be in bed for 7, 10, 14, 20, 21, 28, 30, etc., days compared to the number of people reported to be in bed at days adjacent to them. A similar data clustering problem also exists in the number of physician visits recorded. This problem may be circumvented by suitably grouping the data. Each individual in the sample is classified according to each of the variables: age (1-5, 6-17, 18-34, 35-54, 55-64, 65-above), sex (male and female), economic status (poverty, low, middle, high), and number of bed days (0, 1-2, 3-5, 6-13, 14-above). It should be noted that economic status is based on family income adjusted for family size. Also, the age groupings excluded all individuals under one year of age as they have an inflated number of physician visits due to routine well baby care. Bed day and physician visit groupings were selected so that the cell sizes would be as nearly equal as possible and to compensate for the data clustering problem mentioned above. Next, an individual is classified according to the number of physician visits (0, 1, 2-3, 4-6, 7-above) in addition to the four variables above. Thus, we have four dimensional and five dimensional contingency tables and an analysis is performed using these tables.

Goodman (1970, 1971) has described techniques to analyze multidimensional contingency tables. Related work in this area also include Bishop (1969), Haberman (1974), Fienberg (1970), and Bhapkar and Koch (1968) and several others who are referenced in these papers. We draw upon these papers to analyze the data at hand.

In contingency tables, the terms "response" and "factor" are used when referring to dependent and independent variables, respectively. There are four principal types of multidimensional contingency tables (a) the multiresponse, no factor tables (b) the multiresponse, unifactor tables (c) the multiresponse, multifactor tables (d) the uni-response, multifactor tables. The problems considered in the above mentioned papers are for each of the types of the contingency tables models describing the possible multiplicative interactions and the selection of models that fit the data in the table.

Depending on the type of contingency table under consideration, the term interaction has been given two distinct interpretations. One is related to the nature of the association among responses; the other, to the nature of the way in which

factors combine to determine the responses and functions thereof. In many experimental situations, it is more or less obvious whether a particular dimension is a response or a factor, so that the problem belongs to one of the types. On the other hand, in some instances a particular dimension may be viewed either as a response or a factor in which case the problem can be approached from different points of view.

The health survey data which is arranged in the form of four and five dimensional contingency tables fits into type (d) and (c), respectively. With uni-response and multifactor experiments, we are interested in the way in which the factors combine to determine the response. For exaample, do each of the factors, age, sex, and economic status affect the distribution of the response (bed days)? What order of the interactions between which of the factors affect the response? When the experiment is of multiresponse, multifactor type, then both the relationship among the responses and the way in which the factors combine are of interest. For example, is the dependence of the response (number of physician visits) on the interaction between age and economic status independent of sex? Is the measure of association between the number of bed days and the number of physician visits within each category of the economic status independent of sex and age?

II. MODELS

As this paper deals only with applications, mathematical theory dealing with the formulation of models including estimation of parameters in the model and tests of goodness of fit of models are not discussed here. Interested readers should consult Fienberg (1970) and Goodman (1971). Some comments on models and on selection of models as they relate to data at hand are made.

Once it is decided on the type of contingency table the general problem is one of selecting an unsaturated model. The saturated model is often an overly complicated one and a simple model in which some of the interaction terms are set equal to zero but still providing an adequate fit of the data is preferred. In any higher dimensional contingency table, there is an entire class of models, any one of which might be used to fit the data. One cannot simply test the goodness-of-fit of each model separately, because the test statistics are not independent, and thus the significance levels are not known. There is no best method of model selection and different approaches have been suggested by Bishop (1969), Goodman (1971), and Ku and Kullback (1968). We will adopt a version of partitioning technique combined with stepwise procedures discussed by Goodman (1971) in the analysis of our data.

A nested hierarchy of models in which each of the models contain the previous ones as special cases are formulated. In two successive models, the likelihood ratio statistic for the simple model (H) is partitioned into additive components due to a complex model (H') and to the differences between the two models. The difference in test statistics is asymptotically chi-square with degrees of freedom equal to the difference in the degrees of freedom associated with the two models. In a hierarchy of models, a sequence of tests are done until the difference between successive models is significant or the complex model of the two is insignificant.

In the four-way contingency table the four dimensions pertain to variables: bed days (P), sex (S), economic status (T), and age (U). The variable P is a response variable and so a logit model may be fitted instead of a log-linear model. Instead of working with logit models to select an unsaturated logit model, we can work with those unsaturated log linear models which always have all the terms corresponding to the main affects, interactions of the variables S, T, U. As an illustration consider the following log linear models.

$$\log F_{ijk\ell} = \mu + \lambda^S_j + \lambda^T_k + \lambda^U_\ell + \lambda^{ST}_{jk} + \lambda^{SU}_{j\ell} + \lambda^{TU}_{k\ell} + \lambda^{STU}_{jk\ell} + \lambda^P_i$$

$$\log F_{ijk\ell} = \mu + \lambda^S_j + \lambda^T_k + \lambda^U_\ell + \lambda^{ST}_{jk} + \lambda^{SU}_{j\ell} + \lambda^{TU}_{k\ell} + \lambda^{STU}_{jk\ell} + \lambda^P_i + \lambda^{SP}_{ij}$$

$$\log F_{ijk\ell} = \mu + \lambda^S_j + \lambda^T_k + \lambda^U_\ell + \lambda^{ST}_{jk} + \lambda^{SU}_{j\ell} + \lambda^{TU}_{k\ell} + \lambda^{STU}_{jk\ell} + \lambda^P_i + \lambda^{SP}_{ij} + \lambda^{TP}_{ik} + \lambda^{STP}_{ijk}.$$

In the above, the first model fit hypothesizes that the logits of bed days depend only on the main effects of the grand mean, the second model hypothesizes that they also depend on the main effects of the variable sex, and the third model hypothesizes further dependence on the main effect of the variable economic status and the interaction effect of economic status and sex.

We note here, in a logit situation the λ's with one or more of superscripts

S, T, U corresponding to factors are added first to the model. This is done because the other variables are viewed as fixed, in the same way that the independent variables in multiple regression are viewed as fixed. Thus all effects pertaining to the factors are first removed from the table before trying to fit models to the logits.

III. APPLICATION

A program called C-TAB developed by Haberman (1973) is used for the computations. This program computes expected cell counts, standardized residuals, and estimated effects. Table 1 gives the likelihood ratio (χ^2_L) chi-square values for hypotheses pertaining to models in which the number of bed days is the response variable and age, sex and economic status are the factors.

TABLE 1

Chi-square values for hypotheses pertaining to logit models: Number of bed days (P) is response and sex (S), economic status (T), and age (U) are factors with G=STU.

Hyp	Terms in the model	d.f.	χ^2_L
1	G,P	188	752.2
2	G,PU	168	392.6
3	G,PU,PT	156	272.9
4	G,PT	176	618.1
5	G,PT,PU,PS	152	212.1
6	G,PTU,PS	92	119.4
7	G,PTU	96	179.1
8	G,PS	184	689.9
9	G,PT,PS	172	561.7
10	G,PU,PS	164	327.6
11	G,PST,PU	140	199.5
12	G,PST	160	548.9
13	G,PSU,PT	132	171.3
14	G,PSU	144	288.7
15	G,PTU,PST	80	106.1
16	G,PTU,PSU	72	77.3
17	G,PST,PSU	120	159.9
18	G,PTU,PST,PSU	60	64.7

An interaction term between some variables in any specific model imply all lower order interaction terms between these variables are also present in the model. We note that the third and lower order interaction terms between factors sex, age, and economic status are always present in the model. All possible models are considered in Table 1 but there are practical limitations on computation as dimensions increase. We first use the stepwise procedure to select a model that fits the data adequately. In this procedure each of the λ's present for each new model considered are reexamined. From Table 1, we see that H_1 to H_5 do not fit the data; thus the factors sex, age, and economic status of an individual affect the number of bed days. We now determine which of the three two factor λ's should be added to H_5 in order to improve the fit obtained with H_5. Calculating likelihood ratio chi-square for each of the corresponding three hypotheses, we find that H_6 gives better fit. The significance of adding PTU to H_5 can be seen by calculating the difference between the two likelihood ratio statistics (212.1-119.4=92.7), and judge the magnitude using the chi-square distribution with 152-92-=60 degrees of freedom. The addition of PTU to H_5 has improved the fit significantly at the .05 level. This means the interaction between age and economic status affect the distribution of the number of bed days. The effect of sex on bed days in addition to the effect of interaction between age and economic status is judged by noting the difference 179.1-119.4=59.7 (comparing H_7 with H_6) with 96-92=4 degrees of freedom to be significant at 0.05 level. We now determine which of the other 2 factor λ's should be added to H_6 in order to improve the fit. The addition of PST did not improve the fit significantly over H_6 (compare H_6 with H_{15}); but the addition of PSU to H_6 has improved the fit significantly (compare H_6 with H_{16}). Further, the inclusion of PTU in H_{16} contributed significantly to the fit of H16 (compare H_{13} with H_{16}). Finally the significance of PSU and PTU and insignificance of PST are again noted by comparing H_{15} with H_{18}, H_{17} with H_{18}, and H_{16} with H_{18}, respectively. Thus, model H_{16} fits the data adequately. Table 2 describes the forward selection procedure in which only the most recently entered λ's in the model are examined.

TABLE 2

Analysis of the Logits of Variable P by Forward Selection Method

Source of variation	d.f.	χ^2_L
Total variation of logits of P	188	752.2*
Due to PU	20	359.6*
Due to PT/PU	12	119.7*
Due to PS/PT,PU	4	60.8*
Due to PTU/PT,PU,PS	60	92.7*
Due to PSU/PTU,PT,PU,PS	20	42.1*
Due to PTS/PTU,PSU,PT,PU,PS	12	12.6

* Denotes significance at the 0.05 level.

The two methods have yielded the same solution in our case but this does not mean they will do so always (Draper and Smith; 1966). An interpretation of model H_{16} is; main effects of economic status, age, sex on bed days are significant, interaction effects of economic status and age, age and sex on bed days are signifi-

cant, while the interaction effect of economic status and sex on bed days is not significant. The factors economic status and sex affect the number of bed days independently within the age level.

Similar analysis is carried out using the number of physician visits (V) (0, 1-3, 4-6, 7-above) as the response variable and the number of bed days (0, 1-5, 6-13, 7-above) age (1-17, 18-34, 35-64, 65-above), sex, and economic status are the factors. The pooling of certain classes in variables visits, bed days, and age, is done to avoid certain programming difficulties. The fourth and lower order interaction terms between factors bed days, sex, age, and economic status are always present in the model. The chi-square values are given in Table 3.

TABLE 3

Chi-square values for hypotheses pertaining to logit models: Number of physician visits (V) is response and number of bed days (P), sex (S), economic status (T), and age (U) are factors with H = PSTU.

Hyp	Terms in the model	d.f.	χ^2_L
1	H,VP,VT,VS,VU	351	537.25
2	H,VPT,VS,VU	324	499.17
3	H,VPS,VT,VU	342	520.59
4	H,VPU,VT,VS	324	470.91
5	H,VTS,VP,VU	342	520.90
6	H,VTU,VP,VS	324	454.00
7	H,VSU,VP,VT	342	469.79
8	H,VPT,VSU	315	431.16
9	H,VPS,VTU	315	437.20
10	H,VPU,VTS	315	454.52
11	H,VTU,VSU,VP	315	397.61
12	H,VTU,VTS,VP	315	439.29
13	H,VTU,VPU,VS	297	390.87
14	H,VTU,VPT,VS	297	417.77
15	H,VTU,VSU,VPS	306	382.77
16	H,VTU,VSU,VPT	288	334.81
17	H,VTU,VSU,VPU	288	360.34
18	H,VTU,VSU,VTS,VP	306	385.60
19	H,VTU,VSU,VPU,VPS	279	318.05
20	H,VTU,VSU,VPU,VPT	261	297.84
21	H,VTU,VSU,VPU,VTS	279	322.95
22	H,VTU,VSU,VPU,VPS,VPT	252	282.51
23	H,VTU,VSU,VPU,VPS,VTS	270	307.26
24	H,VTU,VSU,VPU,VPT,VTS	252	285.58
25	H,VPTS,VU	279	437.05
26	H,VPTU,VS	189	268.97
27	H,VTSU,VP	279	337.75
28	H,VPSU,VT	279	367.01

As in Table 1 the stepwise procedure may be used to select a model but as the dimensions increase this involves large number of computations. Hence a forward selection procedure (Table 4) is employed in choosing a model. We see from Table 4 model H_{17} fits the data. The interaction effects of bed days and age, sex and age, and economic status and age on number of physician visits are significant. In comparing this model with the model obtained earlier for the number of bed days as response variable the interaction terms remained the same in both the cases. Additionally the interaction between bed days and age is also present here.

TABLE 4

Analysis of the Logits of Variable V by Forward Selection Method

Source of variation	d.f.	χ^2_L
Total variation of logits of V	381	1926.71*
Due to VP,VT,VS,VU	30	1389.46*
Due to VTU/VP,VT,VS,VU	27	83.25*
Due to VSU/VTU,VP,VT,VS,VU	9	56.39*
Due to VPU/VSU,VTU,VP,VT, VS,VU	27	62.80*
Due to VPT/VPU,VSU,VTU,VP, VT,VS,VU	27	36.97
Due to residual interaction effects	261	297.84

*Denotes significance at the .05 level.

One may also fit a model using both the number of physician visits and the number of bed days as response variables. Tables 5 and 6 show such an analysis.

TABLE 5

Chi-square values for hypotheses pertaining to logit models: Number of physician visits (V) and number of bed days (P) are response variables and sex (S), economic status (T), and age (U) are factors with G=STU.

Hyp	Terms in the model	d.f.	χ^2_L
1	G,VPT,VPS,VPU	360	537.22
2	G,VPTS	360	1118.11
3	G,VPTU	240	420.36
4	G,VPSU	360	600.85
5	G,VPTU,VPS	225	317.51
6	G,VPSU,VPT	315	427.87
7	G,VPTS,VPU	315	467.67
8	G,VPTS,VPTU	180	250.38
9	G,VPTS,VPSU	270	360.67
10	G,VPTU,VPSU	180	222.11
11	G,VPTU,VPSU,VPTS	135	152.23

From Table 6 we find model H_{11} is an adequate fit to the data. The interaction effects of economic status and age, economic status and sex, and age and sex on number of physician visits and number of bed days are significant.

TABLE 6

Analysis of the Logits of Variables V and P by Forward Selection Method

Source of Variation	d.f.	x^2_L
Due to VPTU/VPT,VPS,VPU	135	219.71*
Due to VPSU/VPTU,VPT,VPS VPU	45	95.4*
Due to VPTS/VPSU,VPTU,VPT, VPS,VPU	45	69.88*
Due to other interaction effects on V and P	135	152.23

*Denotes significance at the .05 level.

IV. CONCLUSION

With the increasing number of health surveys, the usefulness of log-linear modelling analysis to the data is explored. The resulting models indicate the interactive factors on measures of use and need of health services. A practical interpretation of interactive factors remains to be done.

In addition, similar household survey data was collected again in 1974 for the same geographical population in Rhode Island. Further research will involve construction of similar models with the 1974 data. The models from the 1972 and 1974 data can then be compared.

FOOTNOTES

*This research was supported in part by Rhode Island Health Services Research, Inc. (SEARCH). The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of SEARCH.

**The SEARCH survey was conducted under a grant from the National Center for Health Services Research (#1-R18-HS-00720).

REFERENCES

1. Bhapkar, V. R. and Koch, G. C. (1968). On the Hypothesis of 'No Interaction' in Contingency Tables. Biometrics 24, 567-596.

2. Bishop, Y. M. M. (1969). Full Contingency tables, logits, and split contingency tables. Biometrics, 25, 383-400.

3. Draper, N. R. and Smith, H. (1966). Applied Regression Analysis. John Wiley and Sons, Inc., New York.

4. Fienberg, S. E. (1970). The Analysis of Multidimensional Contingency Tables. Ecology 51, 419-433.

5. Goodman, L. A. (1970). The Multivariate Analysis of Qualitative Data: Interactions Among Multiple Classifications. J. Amer. Statist. Assoc. 65, 226-256.

6. Goodman, L. A. (1971). The Analysis of Multidimensional Contingency Tables: Stepwise Procedures and Direct Estimation Methods for Building Models for Multiple Classifications. Technometrics 13, 33-61.

7. Haberman, S. J. (1973). C-TAB A FORTRAN IV Program, National Educational Resources, Inc., Ann Arbor, Michigan.

8. Haberman, S. J. (1974). The Analysis of Frequency Data. University of Chicago Press: Chicago.

9. Ku, H. H. and Kullback, S. (1968). Interaction in multidimensional contingency tables: An information theoretic approach. J. Res. National Bureau of Standards 72B, 159-199.

10. Thornberry, O.; Scott, H. D. and Branson, M. H. Methodology of a Health Interview Survey for a Population of One Million. Paper presented at the 101st Annual APHA meeting, San Francisco, November 1973.