DISCRETE MULTIVARIATE ANALYSIS OF SOME ACTUARIAL DATA*

Thomas N. Herzog

Abstract

This paper shows how discrete multivariate analysis (or multidimensional contingency table methods) may be applied to data arising in actuarial work. A brief introduction to the theory is followed by two detailed examples of its application. Finally, the concluding remarks indicate how the procedure could be incorporated into a complete rate-making procedure.

*This work was performed for the U.S. Department of Housing and Urban Development.

We will describe a procedure which is one of a number of methods generally described as (multidimensional) contingency table, or discrete multivariate analysis. These methods have received considerable attention in the recent statistical literature, and a number of specific algorithms for implementing them has been developed and programmed. The recent book of Bishop, Fienberg, and Holland [1] gives an excellent overall account of the field and provides extensive references to the literature. Fienberg [2] provides a less sophisticated treatment of this topic. The specific approach discussed here was developed by Professor Solomon Kullback of The George Washington University (see [3]) and is an information-theoretic approach to contingency table analysis. C. Terrence Ireland directed the implementation of an algorithm for this procedure in both APL and PL/I. Documentation of the algorithm and examples of its application are found in [3 and 4]. Other APL and FORTRAN programs are available from Fox (see [5]) and Goodman (see [6, especially page 468]), respectively.

The paper is divided into four parts. In the first part, we present a brief introduction to the theory of discrete multivariate analysis. In parts II and III, we discuss two applications of the procedure. Finally, in the last part we indicate how this technique could be incorporated into a complete rate-making procedure.

I. THEORY

1. Introduction

Contingency table methods are appropriate for data that have been cross-classified by a number of variables of interest. Each variable is required to be categorical (i.e., discrete), so that each observed data point can be placed into exactly one of a finite number of categories of each variable. The restriction to categorical data is not prohibitive, since data measured on a continuous scale (such as mortgage amount) can be easily categorized (for example, by partitioning the range of mortgage amounts into $5,000 intervals). When all variables are considered simultaneously, they determine a cross-classification into cells. The number of cells, M, is equal to the product of the number of categories in each variable. For example, if there are three variables of interest, with 3, 2, and 5 categories, respectively, then the cross-classification (or contingency table) defined by them has $M = 3 \cdot 2 \cdot 5 = 30$ cells.

The reader should be familiar with two-dimensional tables such as the following table which cross-classifies the income of actuaries by their academic record.

Income of Actuaries versus Academic Record

Academic Record	Income		
	Low	Medium	High
Low	5	17	18
Medium	16	38	26
High	9	15	6

Since there are three income classifications and three academic record classifications, the table consists of 3·3=9 cells. The procedure we are concerned with involves a generalization of the well-known 2-dimensional table to a higher number of dimensions.

In contingency table analysis, the data are assumed to obey an underlying multinomial distribution over all M cells. To each cell is associated the probability that a data point selected at random will be classified in that cell. The two goals of contingency table analysis are: (1) to describe the observed data as simply as possible and (2) to obtain as good a fit as possible. By "a simple description" we mean that we can model the underlying multinomial distribution with relatively few parameters (i.e., main effects and higher order interaction terms). Often the two goals are in conflict because the fit usually increases along with the complexity of the model. Thus, the data analyst must balance the trade-off between the simplicity of the structural model and the closeness of the fit.

The results of the contingency table analysis are:

(1) a model which attempts to maximize both structural simplicity and goodness-of-fit,

(2) estimated frequency counts (under the model) for each cell,

(3) an overall measure of fit of the model,

(4) an analysis of which variables and interactions are present in the model and which are not, and

(5) a measure of the relative importance of the variables and interactions used.

2. Log-linear Structure

The analysis begins with a log-linear model for the underlying multinomial distribution; that is, the logarithms of the cell probabilities are expressed as linear functions of the main effects and interactions of the variables. For purposes of illustration, we consider a data set with 3 variables, I, J, and K. We suppose that variable I has 2 categories (identified as $i = 1$ or 2), variable J has 3 ($j = 1, 2,$ or 3), and variable K has 2 ($k = 1$ or 2). Thus the total number of cells is $M = 2 \cdot 3 \cdot 2 = 12$. Let $p(i,j,k)$ denote the underlying probability that a data point selected at random will be classified in cell (i,j,k). The log-linear model represents each $p(i,j,k)$ in the following form:

$$\log p(i,j,k) = u + a_i^I + a_j^J + a_k^K + a_{ij}^{IJ} + a_{ik}^{IK} + a_{jk}^{JK} + a_{ijk}^{IJK}$$

where i = 1, 2; j = 1, 2, 3; k = 1, 2; and log is the natural logarithm. The 8 terms have the following interpretations:

u is a general (or overall) mean.

a_i^I measures the effect of the variable I alone on the probability p(i,j,k). Since these terms measure deviations from the general mean, they must sum to 0 over all values of i : $a_1^I + a_2^I = 0$. Thus, there is only one independent parameter a_i^I. The a_j^J and a_k^K terms measure similar main effects for the J and K variables.

a_{ij}^{IJ} measures the interaction of variables I and J. Again these measure deviations from the effects of the variables I and J alone, so that $a_{1j}^{IJ} + a_{2j}^{IJ} = 0$ for any j (and similarly for i).

a_{ijk}^{IJK} measures the 3-variable interaction of all 3 variables.

Again, $a_{1jk}^{IJK} + a_{2jk}^{IJK} = 0$ for any pair (j,k).

Thus the complete description of all probabilities p(i,j,k) requires 12 independent parameters:

1 general mean;

4 main effects (1 for I, 2 for J, and 1 for K);

5 2-variable interactions (2 for IJ, 1 for IK, and 2 for JK);

2 3-variable interactions a_{111}^{IJK} and a_{121}^{IJK}.

All of the other parameters can be calculated from these, since their sum across any single variable is 0. Since the data require M = 12 cells, the system is completely determined.

The goal of the analysis may now be expressed in terms of the a-parameters. A structural model for the data is specified by assuming that a certain set of the a-parameters will be used to estimate the probabilities p(i,j,k) (and thus that the main effects and interactions represented by these parameters will be used in the model) and that all other a-parameters will be assumed to be 0 (so that these main effects and interactions are assumed not to exist). If the model fits the data well, then it partitions the main effects and interactions into two groups: those which are important in describing the data and those which are not.

The contingency table models are hierarchical models in the sense that if a particular higher-order interaction term is included in the model, then so are all lower-order interaction terms. For example, if the term a^{IJ}_{ij} appears in the model, then so do a^{I}_{i} and a^{J}_{j}.

Typical models are defined as follows:

(1) The only parameter used is the general mean u. The model says that log p(i,j,k) = u for each p(i,j,k). Thus none of the variables has any effect, and the observed data points can be assumed to be distributed uniformly over the M cells.

(2) Only the general mean and the main effects are used. Now $\log p(i,j,k) = u + a^{I}_{i} + a^{J}_{j} + a^{K}_{k}$: the logarithm of

each cell probability is the sum of the general mean and a main effect from each variable. If both sides of the equation are exponentiated, it is seen that each cell probability is the product of four factors, one being the "general mean" and the other three being the main effects of each of the variables. This is the classical independence model.

(3) The general mean, all main effects, and a single 2-variable interaction effect are used. If the JK interaction is the one present, then log p(i,j,k) $= u + a_i^I + a_j^J + a_k^K + a_{jk}^{JK}$.

The interpretation of this model is that the variable I is independent of the various combinations of J and K.

(4) The general mean, all three main effects, and two 2-variable interactions--IJ and JK--are used.

$$\log p(i,j,k) = u + a_i^I + a_j^J + a_k^K + a_{ij}^{IJ} + a_{jk}^{JK} .$$

This is a conditional independence model: given variable J, variables I and K are independent.

3. <u>The Algorithm</u>

The algorithm CONTAB, which was used to implement the log-linear modeling procedure, employs an iterative fitting scheme. Each model to be fit specifies that certain interactions among the variables are to be included. Using the phrase "marginal total" to mean a sum across one or more variables, we may restate the previous statement as follows:

The specified marginal totals of the fitted table must agree with the corresponding marginal totals of the original table. For instance, in the 3-variable example considered previously, let $x(i,j,k)$ denote the observed count in cell (i,j,k). Thus the IJ marginal is $\sum_k x(i,j,k)$ and is written as $x(i,j,\cdot)$, where the dot replaces the variable which has been summed. Similarly, the K marginal is $\sum_{ij} x(i,j,k) = x(\cdot,\cdot,k)$. In example (3) above the general mean, all main effects, and the JK interaction are to be present in the model. This means that the grand total $N = x(\cdot,\cdot,\cdot)$, all of the 1-variable marginals $x(i,\cdot,\cdot)$, $x(\cdot, j, \cdot)$, and $x(\cdot,\cdot, k)$, and the 2-variable JK marginals $x(\cdot,j,k)$ are fixed by the model to have the same values as in the original data.

The algorithm CONTAB begins with a uniform distribution of N observations over M cells, and successively adjusts the cell entries so that each of the marginals fixed by the model agrees with its specified value. Since the adjustment to fix one marginal will usually destroy the adjustment for the marginals previously considered, the process must cycle repeatedly through all marginals to be fitted. The algorithm does converge* (the individual cell entries each approach a limiting value); the iterative process terminates when a cycle through all marginals changes the cell entries by less than a pre-specified amount.

*See, for example, [1; pages 85-86].

4. Goodness-of-fit

The goodness-of-fit of a model to the observed data is measured by the Information Statistic of Kullback (which is equivalent to the likelihood ratio statistic of [1]). Let $x(i,j,k)$ be the observed count in cell (i,j,k), and let $x^*(i,j,k)$ be the count predicted by a structural model. The Information Statistic, which measures the fit of the predicted distribution x^* to the observed distribution x, is

$$I(x:x^*) = \sum_{ijk} x(i,j,k) \cdot \log \frac{x(i,j,k)}{x^*(i,j,k)}.$$

Under the null hypothesis that the model is correct, the statistic $2I(x:x^*)$ has a distribution which is asymptotically chi-square, with degrees of freedom equal to the number of cells in the table minus the number of independent parameters specified by the model. The value of this statistic can be compared to tables of the chi-square distribution in order to test the goodness-of-fit.

The Information Statistic can be used to test a hierarchical collection of models. Suppose that x_1^* and x_2^* are the distributions predicted from two different structural models, and that all the a-parameters used in the model for x_1^* are also used in the model for x_2^*. Then the statistic $2I(x_2^*:x_1^*)$ satisfies the relation

$$2I(x:x_1^*) = 2I(x:x_2^*) + 2I(x_2^*:x_1^*).$$

The distribution of $2I(x_2^*:x_1^*)$ is also asymptotically chi-square, with degrees of freedom equal to the number of independent parameters specified by x_2^* which are not specified by x_1^*. This statistic may be used to test the hypothesis that the parameters specified by x_2^* but not by x_1^* are statistically significant.

5. <u>An Alternative Measure of Fit: The I^2-Statistic</u>

In the remainder of this paper we will assume that there is one dependent variable (the result to be predicted) and that the other (say n) variables may be regarded as independent (or predictor) variables. We will now define an alternate measure of fit for this type of model. This alternative measure of fit for log-linear models is described by Goodman [7; p. 246] and Scheuren [8; pp. 163-4], where it is referred to as the "I^2 - statistic."

As above, we let x denote the observed distribution of an (n+1)-dimensional table. We also let x_o^* denote the "basic" or "benchmark" distribution (or model) which usually assumes complete independence between the n "independent" variables and the single "dependent" variable. Thus, the basic model x_o^* is formed using only the n-way marginal totals of the n independent variables and the one-way marginal totals of the dependent variable. For each model (or hypothesis), x_1^*, which we consider, we can interpret the quantity

$$I^2 = \frac{I(x:x_o^*) - I(x:x_1^*)}{I(x:x_o^*)}$$

to be the ratio of "explained variation" to "total variation." In other words, I^2 may be considered to be the proportion of variation explained by the addition of those interaction terms in model x_1^* that are not in model x_o^*. Since all of the interaction terms found in x_o^* appear in all alternative models, x_1^*, $I(x:x_o^*) \geq I(x:x_1^*) \geq 0$; hence, $0 \leq I^2 \leq 1$.

6. Odds of Observing Particular Outcomes of the Dependent Variable

Consider the previous example involving the three variables I,J, and K. Suppose that K is the dependent variable and that I and J are independent variables. The variable K has two categories. One convenient way to predict the value of the variable K is to estimate the odds that K=1 compared to K=2 for a given combination of values of I and J. For I=i and J=j, these odds are specified by the estimate of the ratio p(i,j,1)/p(i,j,2). Under the general model, the natural log of this estimated ratio is:

$$\log \frac{p(i,j,1)}{p(i,j,2)} = (a_1^K - a_2^K) + (a_{i1}^{IK} - a_{i2}^{IK})$$

$$+ (a_{j1}^{JK} - a_{j2}^{JK}) + (a_{ij1}^{IJK} - a_{ij2}^{IJK})$$

$$= 2a_1^K + 2a_{i1}^{IK} + 2a_{j1}^{JK} + 2a_{ij1}^{IJK} .$$

$$= 2\left[a_1^K + a_{i1}^{IK} + a_{j1}^{JK} + a_{ij1}^{IJK} \right]$$

Thus, only the interaction terms (or parameters) involving K are needed to specify the predicted odds.

II. FHA SINGLE-FAMILY MORTGAGES ENDORSED IN CALIFORNIA DURING 1974

A. INTRODUCTION

Our first example is based on some data arising from a class of FHA mortgages.

The primary goal of this work is to determine which variables, if any, are useful in predicting the eventual default of such mortgages. The data consist of 19,230 mortgages endorsed in California during calendar year 1974. At the end of calendar year 1978, four hundred and two of these mortgages had already resulted in claim terminations. Thus, the overall claim termination rate was (402 ÷ 19,230) or approximately 2.1 percent.

The 19,230 mortgages were initially assigned to one of 1728 cells of a 5-dimensional (2 by 3 by 6 by 8 by 6) contingency table. The following variables were employed in this table:

1. mortgage status (claim termination or still in force). (2 levels)
2. construction status (new, existing, or HUD acquired property). (3 levels)

3. mortgage amount* ($5,000 - 9,999; $10,000 - 14,999; $15,000 - 19,999; $20,000 - 24,999; $25,000 - 29,999; or ≥ $30,000). (6 levels)

4. loan-to-value ratio** (missing; 0.0 - 89.9 percent; 90.0 - 94.9 percent; 95.0 - 95.9 percent; 96.0 - 96.9 percent; 97.0 - 97.9 percent; 98.0 - 98.9 percent; or 99.0 - 99.9 percent) (8 levels)

5. office (Los Angeles, San Francisco, Sacramento, San Diego, Fresno, Santa Ana). (6 levels)

We first cross-classified each of the four predictor variables with mortgage status, producing the following two-way tables of observed frequency counts:

Mortgage Status versus Construction Status

	Construction Status		
	New	Existing	HUD Acquired
Number of claim terminations	11	294	97
Original number of mortgages written	4,245	13,155	1,830
Claim termination rate	0.003	0.022	0.053

* There were no mortgages written for amounts under $5,000, or over $45,000.

**This is the ratio of the loan amount to the estimated property value.

Mortgage Status versus Mortgage Amount

	Mortgage Amount (in dollars)					
	5,000-9,999	10,000-14,999	15,000-19,999	20,000-24,999	25,000-29,999	≥ 30,000
Number of claim terminations	0	72	140	141	34	15
Original number of mortgages written	54	863	3,270	7,024	5,285	2,734
Claim termination rate	0	0.083	0.043	0.020	0.006	0.005

Mortgage Status versus Loan-to-Value Ratio

	Loan-to-Value Ratio (in percent)							
	< 90.0	90.0-94.9	95.0-95.9	96.0-96.9	97.0-97.9	98.0-98.9	99.0-99.9	Total Non-Missing	Missing
Number of claim terminations	6	61	26	28	30	6	9	166	236
Original number of mortgages written	2,553	5,444	996	688	316	175	114	9,874	8,944
Claim termination rate	0.002	0.011	0.026	0.041	0.095	0.034	0.079	0.017	0.026

Mortgage Status versus Office

	Office					
	Los Angeles	San Francisco	Sacramento	San Diego	Fresno	Santa Ana
Number of claim terminations	189	23	22	25	14	129
Original number of mortgages	4,018	3,548	2,968	2,603	1,944	4,149
Claim termination rate	0.047	0.006	0.007	0.010	0.007	0.031

There are a number of important things to note from these four tables. First, there were virtually no claim terminations on new home mortgages. (We observed only 11 claim terminations out of 4,245 mortgages written.) Of the 54 mortgages written for amounts under $10,000, there were no claim terminations. Thus, we can restrict our attention to those mortgages written for existing or acquired property having a mortgage amount of at least $10,000.

As anticipated, the claim termination rate generally tends to increase as the loan-to-value ratio increases; however, the claim termination rate of mortgages whose loan-to-value ratio was missing was substantially higher (2.6 percent versus 1.7 percent) than for those whose loan-to-value ratio was available. This is a potentially serious problem. Because of this bias, it is probably advisable to use special caution when using the loan-to-value ratio variable in subsequent analysis of these mortgages.

Finally, the claim termination rates were substantially higher for the Los Angeles and Santa Ana offices than for the other four California offices.

In the rest of this section of the paper we will discuss our attempts to use contingency-table analysis methods to predict mortgage status.

B. ANALYSIS OF EXISTING MORTGAGE DATA

Since only two of the 97 observed defaults on acquired property mortgages had a loan-to-value ratio present, we decided to restrict our attention to mortgages on existing homes. We began by constructing a 4-dimensional contingency table in which we cross-classified mortgage status (2 levels) by office (6 levels) by loan-to-value ratio missingness status (2 levels) (either missing or present) by mortgage amount (5 levels). Since there are no claim terminations under $10,000, we restricted attention to the five largest mortgage-amount classes.

We first computed the overall claim termination rates. For mortgages with the loan-to-value ratio present the claim termination rate was $(156/7726) \doteq 2.0$ percent;

the others had a claim termination rate of (138/5429)$\doteq$2.5 percent, or a rate somewhat higher than 2.0 percent. So, we next decided to test the (null) hypothesis that mortgage status is (conditionally) independent of the absence or presence of the loan-to-value ratio given the other two variables (office and mortgage amount). We observed a value of 37.2 for the appropriate test statistic. Since this statistic is asymptotically chi-square with 30 degrees of freedom, we were unable to reject the null hypothesis at the 10 percent level of significance. As a result we decided to include the loan-to-value ratio as one of our model's independent variables and thereby to restrict attention to those mortgages for which a loan-to-value ratio was present.

The next step was to construct a (2 by 6 by 3 by 4) contingency table. Here the first two dimensions are as in the last table and the last dimension is based on the mortgage amount, the two highest mortgage amount classes being combined to reduce the number of such classes from 5 to 4. The remaining variable is the loan-to-value ratio which is partitioned as follows:

(0.0-94.9%), (95.0-96.9%) and (97.0-99.9%).

A transposed version of the resulting contingency table constitutes Table I of the appendix. Contingency-table analysis resulted in the following table of values:

Model	Information Statistic times 2	I^2-Value	Degrees of Freedom	Probability of a Larger Value
Constant term only	254.1	-	71	0.00
O	202.2	0.204	66	0.00
L	161.2	0.366	69	0.00
M	154.9	0.390	68	0.00
OL	97.4	0.617	54	0.00
OM	83.0	0.673	48	0.00
LM	106.9	0.579	60	0.00
OL,M	55.0	0.784	51	0.33
OM,L	47.4	0.813	46	0.41
LM,O	58.6	0.769	55	0.34
OL,OM	34.6	0.864	36	0.53
OL,LM	45.9	0.819	45	0.43
OM,LM	38.4	0.849	40	0.54
OL,OM,ML	23.4	0.908	30	0.80

where

O represents office

L represents loan-to-value ratio, and

M represents mortgage amount

For example, the OL model is the log-linear model which includes the office - loan-to-value ratio two-way interaction term.

Since we are only considering hierarchical models, the OL model also contains the office and loan-to-value ratio main effect terms.

The first model consisted only of the constant term. In this model the probability of a claim termination is assumed to be independent of the loan-to-value ratio, mortgage amount, and field office. The resulting information statistic times two is equal to 254.1--a fairly large value for a chi-square distribution with 71 degrees of freedom.

From the last table, we also see that the mortgage amount, M, is the single variable which explains the most variation, i.e., 39.0 percent. The loan-to-value ratio, L, and the office*, O, variables explain 36.6 percent and 20.4 percent respectively. The OM model involving the office-mortgage amount two-way interaction term explains the most variation among the class of models having a single two-way interaction term; specifically, this I^2-value is 67.3 percent. We found that the model involving the mortgage amount - loan-to-value ratio two-way interaction

* The office variable should be considered to be a proxy for location and/or office procedure variables.

term was inferior to the other two models of this class, both of which included the office variable. Thus, it appears that the mortgage amount and the loan-to-value ratio are "explaining a lot of the same variation" compared to the office.

We also considered the three models with one main effect and one two-way interaction term. The model with the office-mortgage amount interaction term appeared to be the best of these.

We next examined the models having a pair of two-way interaction terms. There does not appear to be a standout among these models. Although the OL,OM model has the highest I^2-value (specifically, 0.864), it also has the smallest number of degrees of freedom (36 versus 45 for the OL,LM model).

Finally, we examined the model having three two-way interaction terms. This model has an I^2-value of 0.908 based on 30 degrees of freedom.

It is admittedly difficult to choose among these models. In our appendix, we have presented the predicted cell frequencies for the OL,OM and the OL,OM,ML models. These had the two highest I^2-values.

C. WARNING ! !

The results of this contingency-table analysis should be treated with caution because 31 of the 144 observed cell frequencies were zero and 42 others were less than or equal to 5. In fact, only 7 of the 72 claim termination cells contained more than 5 claim terminations. Thus, the vast majority of the claim termination cells will have expected cell frequencies which are not substantially greater than 5.

III. AUTOMOBILE ACCIDENT DATA

The second example is a manufactured example which has been constructed to demonstrate a potential use of contingency analysis in the automobile insurance risk classification process. The basic data constitute Table VI of the appendix.

Our initial assumption is that there is one dependent variable and three predictor variables. The dependent variable is the accident indicator variable which indicates whether or not an individual was involved in at least one automobile accident during the most recent calendar year. The independent variables are sex, age (which has been partitioned into four categories), and location (which has been partitioned into three categories).

Our goal is to determine which of the sex, age, and location variables are most useful in predicting the value of the accident indicator variable. We have considered a number of models and have summarized the results in Table VII.

The first model consisted of the constant term only. Here the probability of at least one accident was constant over all combinations of sex, age, and location. For this model, the information statistic times two is equal to 413.3--a rather large value for a chi-square distribution with only

23 degrees of freedom. From the next three models, we see that location is the single variable which explains the most variation--62 percent--compared to 4 percent and 36 percent for the sex and age variables, respectively. Of the models containing two main effect terms, the one containing age and location is clearly the best, explaining 87 percent of the variation.

For the models containing a single two-way interaction term, the age-location model stands out with an I^2-value of 93 percent. Yet, even the information statistic corresponding to this model is quite high; i.e., it is still statistically significant at the 0.02 level.

The next group of models considered consists of one two-way interaction term and a single main effect term. Here again the model containing the age-location interaction is the best. This model is a legitimate candidate for the best overall model because the value of its information statistic is relatively low and its I^2-value is quite high.

The last group of models considered each contained a pair of two-way interaction terms. The best one here consisted of the sex-age and age-location terms. This model is also a candidate for the best overall model because the product of two and its information statistic is less than 8, the number of degrees of freedom. It does, however, contain three more parameters than the AL,S model and so may not be preferred if a more parsimonious model is desired.

In Table VIII, we present the actual parameter estimates for the model containing the age-location two-way interaction term and the sex main effect term. Using these estimates we find that, under the AL,S model, the log odds of a 20-year-old urban male having at least one accident are

$$\log \frac{p(1,1,1,1)}{p(2,1,1,1)} = 2[a_1^I + a_{11}^{IS} + a_{11}^{IA} + a_{11}^{IL} + a_{111}^{IAL}]$$

$$= 2[-1.089 + 0.031 + 0.126 + 0.208 + 0.024]$$

$$= 2[-0.7] = -1.4.$$

This corresponds to a probability of

$$\frac{1}{1 + \exp(1.4)} = 0.20$$

of having at least one accident.

IV. CONCLUDING REMARKS

I envision this procedure being used to help compute net premiums in the following manner. First, using the procedure just described, the variables (i.e., main effects and interaction terms) explaining a substantial portion of the total variation are identified and the corresponding estimated cell frequencies are produced. These yield estimated accident probabilities for each combination of the independent (or predictor) variables. Then using regression analysis, or an alternative procedure, a separate expected loss per accident is estimated for each combination. Multiplying each such expected loss by the corresponding accident probability, we obtain an initial estimate of the net premium for each combination of predictor variables. If desired, these estimates may be smoothed further, for example, by using empirical Bayes methods as discussed in Morris and van Slyke [9].

References:

1. Bishop, Y. M.; Fienberg, S. E.; and Holland, P. W.: Discrete Multivariate Analysis: Theory and Practice, Cambridge, Mass., MIT Press, 1975.

2. Fienberg, S. E., The Analysis of Cross-Classified Categorical Data, Cambridge, Mass., MIT Press, 1977.

3. Gokhale, D. V. and Kullback, S.; The Information in Contingency Tables, Dekker, New York, 1978.

4. Fisher, M.; Ireland, C.; Keegel, J.; Kullback, S.; Nolan, J.; and Scheuren, F., Computer Programs on Contingency Table Analysis, Department of Statistics, The George Washington University, Washington, D. C., 1975.

5. Fox, J.; "TAB: An APL Workspace for the Log-Linear Analysis of Contingency Tables," The American Statistician, Vol. 33, No. 3, 1979.

6. Goodman, L. A., Analyzing Qualitative/Categorical Data, Abt Books, Cambridge, Mass., 1978.

7. Goodman, L. A.; "The multivariate analysis of qualitative data: interactions among multiple classifications," Journal of the American Statistical Association, Vol. 65, pp. 226-256, 1970.

8. Scheuren, F. J., "Ransacking CPS tabulations: Applications of the log-linear model to poverty statistics," Annals of Economic and Social Measurement, Vol. 2, No. 2, 1973.

9. Morris, C. and van Slyke, O., "Empirical Bayes Methods for Pricing Insurance Classes," Proceedings of the 1978 American Statistical Association Meeting, Business Statistics Section.

Table I -- Part 1

Observed Number of Existing Mortgages Still in Force by Loan-to-Value Ratio, Office, and Mortgage Amount

MORTGAGE AMOUNT (in dollars)

Loan-to-Value Ratio	Office	10,000-14,999	15,000-19,999	20,000-24,999	≥25,000
<94.9%	Los Angeles	25	116	734	553
	San Francisco	3	36	397	809
	Sacramento	24	48	174	177
	San Diego	3	26	210	585
	Fresno	15	160	203	82
	Santa Ana	38	80	445	583
95.0-96.9%	Los Angeles	11	223	228	61
	San Francisco	2	12	52	49
	Sacramento	19	62	67	33
	San Diego	0	16	50	110
	Fresno	32	124	65	11
	Santa Ana	21	81	86	99
97.0-99.9%	Los Angeles	49	40	15	10
	San Francisco	0	4	11	12
	Sacramento	10	31	20	13
	San Diego	4	9	27	43
	Fresno	15	62	10	2
	Santa Ana	31	58	28	26

Table 1 -- Part 2

Observed Number of Claim Terminations Among Existing Mortgages by Loan-to-Value Ratio, Office, and Mortgage Amount

Loan-to-Value Ratio	Office	MORTGAGE AMOUNT (in dollars) 10,000-14,999	15,000-19,999	20,000-24,999	≥25,000
<94.9%	Los Angeles	1	2	20	9
	San Francisco	0	0	5	2
	Sacramento	0	0	1	0
	San Diego	1	0	3	4
	Fresno	0	1	1	0
	Santa Ana	0	2	7	1
95.0-96.9%	Los Angeles	5	20	7	0
	San Francisco	0	2	5	0
	Sacramento	1	1	0	0
	San Diego	0	0	3	0
	Fresno	0	2	0	0
	Santa Ana	1	2	4	1
97.0-99.9%	Los Angeles	13	1	2	0
	San Francisco	0	1	1	0
	Sacramento	0	1	0	0
	San Diego	1	0	1	1
	Fresno	0	2	1	0
	Santa Ana	5	9	1	2

Table II -- Part 1

Predicted Number* of Existing Mortgages Still in Force by Loan-to-Value Ratio, Office, and Mortgage Amount Under (OL, OM)-Model

MORTGAGE AMOUNT (in dollars)

Loan-to-Value Ratio	Office	10,000-14,999	15,000-19,999	20,000-24,999	≥25,000
<94.9%	Los Angeles	22.9	113.9	736.4	554.8
	San Francisco	3.0	35.3	396.9	809.7
	Sacramento	23.8	47.7	174.6	177.0
	San Diego	3.4	26.0	208.9	585.7
	Fresno	15.0	160.0	203.0	82.0
	Santa Ana	37.4	80.7	446.1	581.8
95.0-96.9%	Los Angeles	12.6	226.7	224.3	59.5
	San Francisco	2.0	12.3	52.2	48.4
	Sacramento	19.5	61.9	66.6	33.0
	San Diego	0.0	16.0	51.2	108.8
	Fresno	32.0	124.6	64.4	11.0
	Santa Ana	21.1	79.7	87.1	99.1
97.0-99.9%	Los Angeles	49.5	38.4	16.3	9.8
	San Francisco	0.0	4.3	10.8	11.8
	Sacramento	9.7	31.4	19.9	13.0
	San Diego	3.6	9.0	26.9	43.5
	Fresno	15.0	61.4	10.6	2.0
	Santa Ana	31.5	58.5	25.9	27.1

*Rounded to one decimal place.

Table II -- Part 2

Predicted Number* of Claim Terminations Among Existing Mortgages by Loan-to-Value Ratio, Office, and Mortgage Amount Under (OL, OM)-Model

		MORTGAGE AMOUNT (in dollars)			
Loan-to-Value Ratio	Office	10,000-14,999	15,000-19,999	20,000-24,999	≥25,000
<94.9%	Los Angeles	3.1	4.1	17.6	7.2
	San Francisco	0.0	0.7	5.1	1.3
	Sacramento	0.2	0.3	0.4	0.0
	San Diego	0.6	0.0	4.1	3.3
	Fresno	0.0	1.0	1.0	0.0
	Santa Ana	0.6	1.3	5.9	2.2
95.0-96.9%	Los Angeles	3.4	16.3	10.7	1.5
	San Francisco	0.0	1.7	4.8	0.6
	Sacramento	0.5	1.1	0.4	0.0
	San Diego	0.0	0.0	1.8	1.2
	Fresno	0.0	1.4	0.6	0.0
	Santa Ana	0.9	3.3	2.9	0.9
97.0-99.9%	Los Angeles	12.5	2.6	0.7	0.2
	San Francisco	0.0	0.7	1.2	0.2
	Sacramento	0.3	0.6	0.1	0.0
	San Diego	1.4	0.0	1.1	0.5
	Fresno	0.0	2.6	0.4	0.0
	Santa Ana	4.5	8.5	3.1	0.9

*Rounded to one decimal place.

Table III -- Part 1

Predicted Number* of Existing Mortgages Still in Force by Loan-to-Value Ratio, Office, and Mortgage Amount Under (OL, OM, ML)-Model

Loan-to-Value Ratio	Office	10,000-14,999	15,000-19,999	20,000-24,999	≥25,000
		MORTGAGE AMOUNT (in dollars)			
<94.9%	Los Angeles	24.7	115.3	734.3	553.7
	San Francisco	3.0	35.6	396.9	809.5
	Sacramento	23.9	47.7	174.4	177.0
	San Diego	3.6	26.0	209.2	585.2
	Fresno	15.0	160.2	202.8	82.0
	Santa Ana	37.8	81.1	445.4	581.7
95.0-96.9%	Los Angeles	11.0	224.9	226.4	60.7
	San Francisco	2.0	11.9	52.2	48.9
	Sacramento	19.4	61.9	66.7	33.0
	San Diego	0.0	16.0	50.4	109.6
	Fresno	32.0	124.4	64.6	11.0
	Santa Ana	20.6	78.9	87.6	99.8
97.0-99.9%	Los Angeles	49.3	38.8	16.3	9.6
	San Francisco	0.0	4.4	10.9	11.7
	Sacramento	9.7	31.4	19.9	13.0
	San Diego	3.4	9.0	27.3	43.3
	Fresno	15.0	61.4	10.6	2.0
	Santa Ana	31.6	59.0	26.0	26.5

*Rounded to one decimal place.

Table III -- Part 2

Predicted Number* of Claim Terminations Among Existing Mortgages by Loan-to-Value Ratio, Office, and Mortgage Amount Under (OL, OM, ML)-Model

		MORTGAGE AMOUNT (in dollars)			
Loan-to-Value Ratio	Office	10,000-14,999	15,000-19,999	20,000-24,999	≥25,000
<94.9%	Los Angeles	1.3	2.7	19.7	8.3
	San Francisco	0.0	0.4	5.1	1.5
	Sacramento	0.1	0.3	0.6	0.0
	San Diego	0.4	0.0	3.8	3.8
	Fresno	0.0	0.8	1.2	0.0
	Santa Ana	0.2	0.9	6.6	2.3
95.0-96.9%	Los Angeles	5.0	18.1	8.6	0.3
	San Francisco	0.0	2.1	4.8	0.1
	Sacramento	0.6	1.1	0.3	0.0
	San Diego	0.0	0.0	2.6	0.4
	Fresno	0.0	1.6	0.4	0.0
	Santa Ana	1.4	4.1	2.4	0.2
97.0-99.9%	Los Angeles	12.7	2.2	0.7	0.4
	San Francisco	0.0	0.6	1.1	0.3
	Sacramento	0.3	0.6	0.1	0.0
	San Diego	1.6	0.0	0.7	0.7
	Fresno	0.0	2.6	0.4	0.0
	Santa Ana	4.4	8.0	3.0	1.5

*Rounded to one decimal place.

Table IV

Predicted Relative Frequency* of Claim Termination Among Existing Mortgages by Loan-to-Value Ratio, Office, and Mortgage Amount Under (OL, OM)-Model

		MORTGAGE AMOUNT (in dollars)			
Loan-to-Value Ratio	Office	10,000-14,999	15,000-19,999	20,000-24,999	≥25,000
<94.9%	Los Angeles	.120	.035	.023	.013
	San Francisco	.000	.018	.013	.002
	Sacramento	.010	.007	.002	.000
	San Diego	.154	.000	.019	.006
	Fresno	.000	.006	.005	.000
	Santa Ana	.016	.016	.013	.004
95.0-96.9%	Los Angeles	.214	.067	.046	.025
	San Francisco	.000	.119	.084	.011
	Sacramento	.025	.017	.006	.000
	San Diego	.253	.000	.035	.010
	Fresno	.000	.011	.009	.000
	Santa Ana	.039	.039	.033	.009
97.0-99.9%	Los Angeles	.201	.063	.042	.024
	San Francisco	.000	.136	.097	.013
	Sacramento	.027	.019	.007	.000
	San Diego	.277	.000	.039	.012
	Fresno	.000	.041	.034	.000
	Santa Ana	.126	.126	.107	.032

*Rounded to three decimal places.

Table V

Predicted Relative Frequency* of Claim Termination Among Existing Mortgages by Loan-to-Value Ratio, Office, and Mortgage Amount Under (OL, OM, ML)-Model

MORTGAGE AMOUNT (in dollars)

Loan-to-Value Ratio	Office	10,000-14,999	15,000-19,999	20,000-24,999	≥25,000
<94.9%	Los Angeles	.049	.023	.026	.015
	San Francisco	.000	.010	.013	.002
	Sacramento	.004	.006	.003	.000
	San Diego	.099	.000	.018	.007
	Fresno	.000	.005	.006	.000
	Santa Ana	.006	.011	.015	.004
95.0-96.9%	Los Angeles	.313	.075	.036	.004
	San Francisco	.000	.149	.084	.003
	Sacramento	.031	.018	.004	.000
	San Diego	.660	.000	.048	.004
	Fresno	.000	.013	.006	.000
	Santa Ana	.062	.049	.027	.002
97.0-99.9%	Los Angeles	.205	.053	.041	.042
	San Francisco	.000	.110	.093	.027
	Sacramento	.027	.019	.007	.000
	San Diego	.321	.000	.024	.016
	Fresno	.000	.041	.033	.000
	Santa Ana	.123	.120	.104	.055

*Rounded to three decimal places.

Table VI

Number of Insureds Having No Automobile Accidents During 1977

	Sex					
	Male			Female		
	Location			Location		
Age	Urban	Suburban	Rural	Urban	Suburban	Rural
16-24	1600	850	460	1400	780	465
25-39	1250	700	470	1300	670	375
40-64	1050	750	940	1030	750	925
≥ 65	860	900	280	1350	1380	455

Number of Insureds Having At Least One Automobile Accident During 1977

	Sex					
	Male			Female		
	Location			Location		
Age	Urban	Suburban	Rural	Urban	Suburban	Rural
16-24	400	150	40	300	120	35
25-39	250	100	30	200	80	25
40-64	150	50	60	170	50	75
≥ 65	140	100	20	160	120	45

Table VII

Model	Information Statistic times 2	I^2-value	D. F.	Probability of a Larger Value
Constant term only	413.3	----	23	0.00
S	397.3	0.04	22	0.00
A	265.9	0.36	20	0.00
L	157.2	0.62	21	0.00
S,A	255.8	0.38	19	0.00
S,L	142.1	0.66	20	0.00
A,L	54.4	0.87	18	0.00
S,A,L	44.6	0.89	17	0.00
SA	247.3	0.40	16	0.00
SL	133.8	0.68	18	0.00
AL	26.9	0.93	12	0.01
SA,L	56.6	0.86	14	0.00
SL,A	37.1	0.91	15	0.00
AL,S	17.3	0.96	11	0.10
SA,SL	30.5	0.93	12	0.00
SA,AL	7.8	0.98	8	0.45
SL,AL	11.5	0.97	9	0.24

Table VIII

Parameter Estimates of the AL,S Model

$a_{1}^{I} = -1.089$

$a_{11}^{IS} = 0.031$

$a_{11}^{IA} = 0.126$	$a_{12}^{IA} = 0.004$
$a_{11}^{IL} = 0.208$	$a_{12}^{IL} = 0.011$
$a_{111}^{IAL} = 0.024$	$a_{112}^{IAL} = 0.072$
$a_{121}^{IAL} = 0.009$	$a_{122}^{IAL} = 0.079$
$a_{131}^{IAL} = 0.045$	$a_{132}^{IAL} = 0.118$

where I is the accident indicator variable, S = sex, A = age, and L = location.