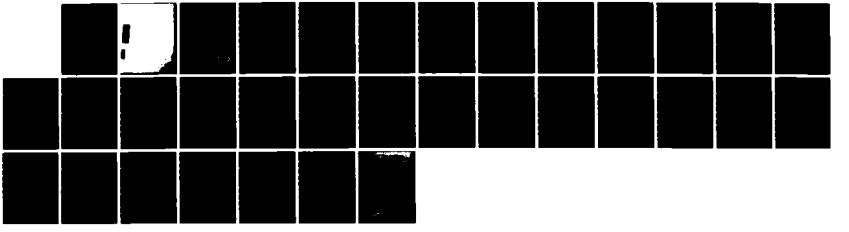
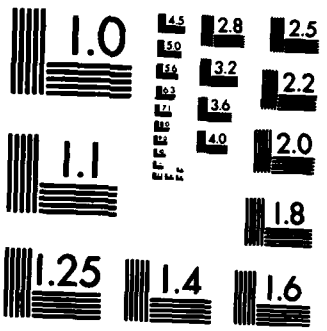
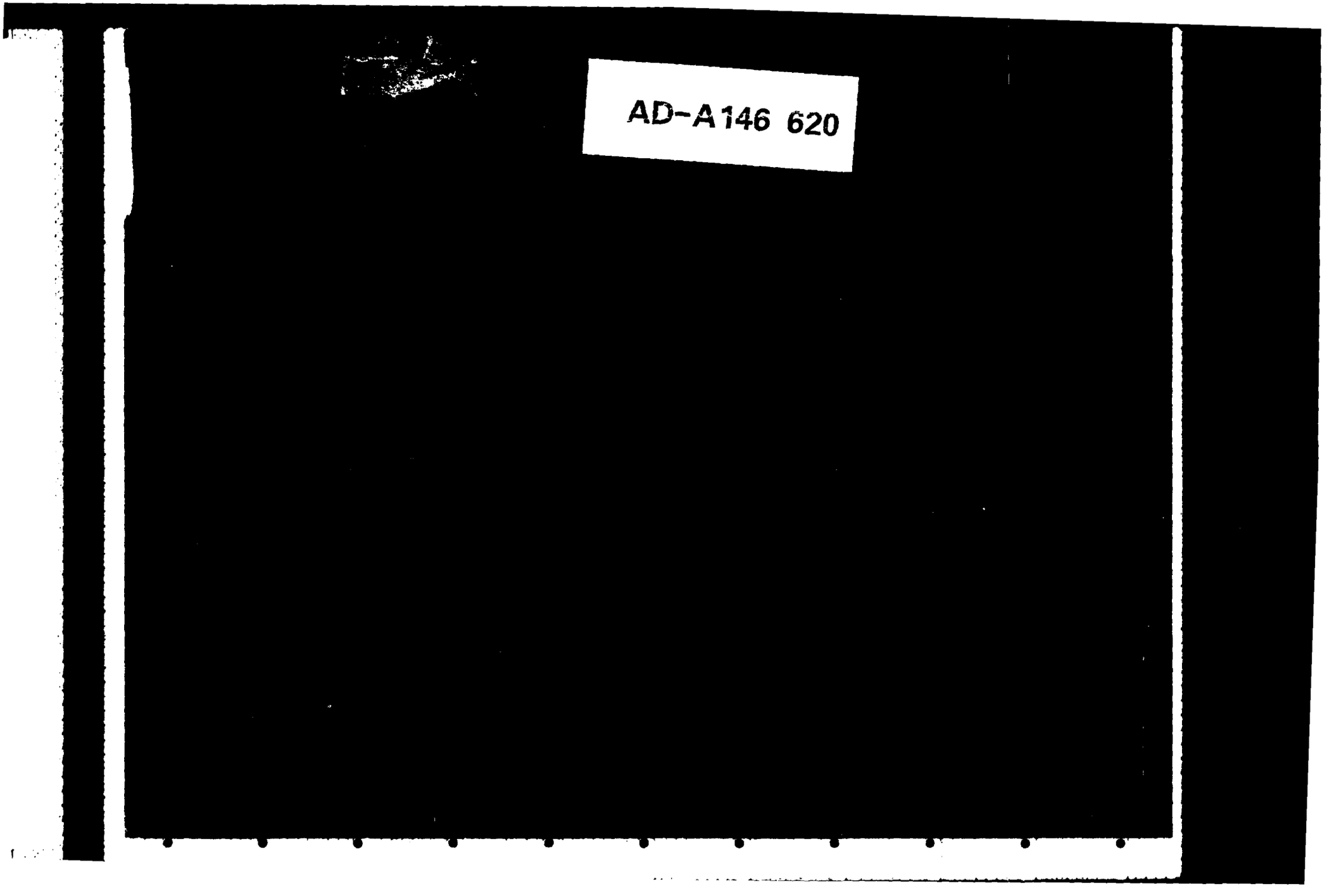
AD-A146 620 LOG LINEAR MODEL APPLICATIONS(U) STANFORD UNIV CA DEPT 1/1
OF STATISTICS H SOLOMON 30 AUG 84 TR-349
N00014-76-C-0475

UNCLASSIFIED F/G 12/1 NL





COPY RESOLUTION TEST CHART



LOG LINEAR MODEL APPLICATIONS

BY

HERBERT SOLOMON

TECHNICAL REPORT NO. 349
AUGUST 30, 1984

Prepared Under Contract
N00014-76-C-0475 (NR-042-267)
For the Office of Naval Research

Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

LOG LINEAR MODEL APPLICATIONS

Herbert Solomon
Stanford University

Methods for multivariate data analysis have been developed and received wide usage in the past 30 years. The advent of the computer in the early 1950's made possible an acceleration in the number of multivariate studies. The first to profit was factor analysis which became a commonplace in exploratory studies. Subsequently, linear discriminant analysis, classification methods and then cluster analyses became part of the emerging excitement in attempting and doing studies either not considered possible before or not thought about at all before because of the lack of computer resources. Another multivariate method has also profited from the computer era. It is multidimensional contingency table analysis through use of the log-linear model. Very recently a two part survey and discussion of the literature of log-linear and logistic categorical data modeling appeared in two issues of the International Statistical Review authored by Imrey, Koch, and Stokes [5] and [6].

In this exposition we will look at some applications of the log-linear model, or as it sometimes is called, logistic response analysis. It may be instructive to view one situation explored about 25 years ago through classification techniques before computer capability to handle the log-linear model and software was available, and then analyze the same data base by contingency table analysis. After this comparison, we will discuss other and more recent direct applications of the log-linear model.

This report is based on an invited talk delivered at the First Pacific Area Statistical Conference, "Recent Developments in Statistical Theory and Data Analysis" held in Tokyo, Japan in December 1982. The Conference was sponsored by the Pacific Statistical Institute and endorsed by the Institute of Statistical Mathematics (Tokyo), the American Statistical Association, and the Statistical Society of Australia.

Classification and Log-Linear Models.

In attempt to examine classification procedures for a set of dichotomous variables for which interactions of higher order were not assumed to be zero, that is, conditions that could violate the usual normality assumptions in discriminant analysis and classification techniques available in the 1950's, Solomon [8] reports on some conclusions for a data base on attitudes toward science elicited from 2982 high school seniors in New Jersey. This data base is then re-examined by Gokhale and Kullback [3] employing a log-linear model. Six items on a 1957 questionnaire "Attitudes Toward Science and Scientific Careers" were chosen on the basis of good discrimination between high and low IQ as measured on a brief IQ vocabulary test. This led to two groups each containing 1491 students. The scoring was such that a student either 'agreed', recorded as 1, or 'disagreed', recorded as 0, with the items listed below

x_1 The development of new ideas is the scientist's greatest source of satisfaction.

x_2 Scientists and engineers should be eliminated from the military draft.

x_3 The scientist will make his maximum contribution to society when he has freedom to work on problems which interest him.

x_4 The monetary compensation of a Nobel Prize winner in physics should be at least equal to that given popular entertainers.

x_5 The free flow of scientific information among scientists is essential to scientific progress.

x_6 The neglect of basic scientific research would be the equivalent of "killing the goose that laid the golden eggs."

For some of the proposed analyses, even the use of six items proved to be cumbersome for the computing recources available at Columbia University about 1958 so that only the first four of the six items above were used in several studies. The frequency distributions for the four items are given in Table 1.

One of the purposes of the study was to contrast the effectiveness of the sum of item responses with the use of the total response vector, that is allow, for example, for the added information in employing the vector (1101) in a classification procedure rather than just the score value equal to three. For classification procedures, other aggregates of information to represent attitudes may be developed that fall between the total response vector and the sum that may give less information than the former and more than the latter. One way to quantify this is to exploit a representation of the joint distribution of responses for n dichotomous items developed by Bahadur [1].

Let X denote the set of all points $x = (x_1,x_2,\ldots,x_n)$, each $x_i = 0$ or 1 and let $p(x)$ be a given probability distribution on X. For each $i = 1,2,\ldots,n$, let $\alpha_i = \Pr\{x_i=1\} = E(x_i)$, that is, the α_i represent the marginal frequencies of the x_i's. Label $p_{[1]}(x_1,x_2,\ldots,x_n)$ the joint probability distribution of the x_i when the x_i's are independently distributed and they have the same marginal distributions as under the given $p(x_1,x_2,\ldots,x_n)$. Then we may write

$$p_{[1]}(x_1,x_2,\ldots,x_n) = \prod_{i=1}^{n} \alpha_i^{x_i}(1-\alpha_i)^{1-x_i}$$

$$p(x) = p_{[1]}(x) \cdot f(x)$$

TABLE 1

JOINT DISTRIBUTION OF RESPONSES ON ATTITUDE TO SCIENCE

x_1, x_2, x_3, x_4	Low IQ		High IQ	
	Frequency	Relative frequency	Frequency	Relative frequency
1111	62	.042	122	.082
1110	70	.047	68	.046
1101	31	.021	33	.022
1100	41	.027	25	.017
1011	283	.190	329	.221
1010	253	.170	247	.166
1001	200	.134	172	.115
1000	305	.205	217	.146
0111	14	.009	20	.013
0110	11	.007	10	.007
0101	11	.007	11	.007
0100	14	.009	9	.006
0011	31	.021	56	.038
0010	46	.031	55	.037
0001	37	.025	64	.043
0000	82	.055	53	.036
Totals	1491	1.000	1491	1.002

where

$$f(x) = 1 + \sum_{i<j} r_{ij} z_i z_j + \sum_{i<j<k} r_{ijk} z_i z_j z_k + \cdots + r_{123\cdots n} z_1 z_2 z_3 \cdots z_n$$

and

$$z_i = \frac{x_i - \alpha_i}{\sqrt{\alpha_i(1-\alpha_i)}}$$

$$r_{ij} = E(z_i z_j)$$

$$r_{ijk} = E(z_i z_j z_k)$$

.....

$$r_{123\cdots n} = E(z_1 z_2 z_3 \cdots z_n) .$$

By dropping out terms, we get approximations to the actual frequencies $p(x)$ and thus arrive at quantitative ways to register the amount of information less than the total response vector or more than the sum. For example, if joint normal distributions prevail, then all correlations higher than second order, r_{ijk}, $r_{ijk\ell}$, etc., are zero; or we could in our four item case assume that some r_{ijk} and $r_{ijk\ell}$ are not zero, an event also tolerated by a log-linear model.

For our four item situation, the following values were obtained

	r_{12}	r_{13}	r_{14}	r_{23}	r_{24}	r_{34}	r_{123}	r_{124}	r_{134}	r_{234}	r_{1234}
High IQ	.003	.143	.111	.180	.020	.140	-.010	-.064	-.041	-.033	.003
Low IQ	.049	.144	.043	.155	.096	.125	-.006	-.012	.002	.002	.002

and

	α_1	α_2	α_3	α_4
High IQ	.821	.159	.505	.436
Low IQ	.801	.189	.599	.530

The main thrust of this study is to provide a procedure for classifying a student into the high IQ or low IQ categories based on a representation of the joint distribution (actual or approximated) of his or her four or six responses. Later, in log-linear model language, we can discuss the odds ratios for these events or equivalently the probability of being placed in one category or the other given a response vector.

We now use a likelihood ratio procedure for classifying an individual in High IQ or Low IQ, i.e., compute

$$L(x) = \frac{h(x)}{\ell(x)}$$

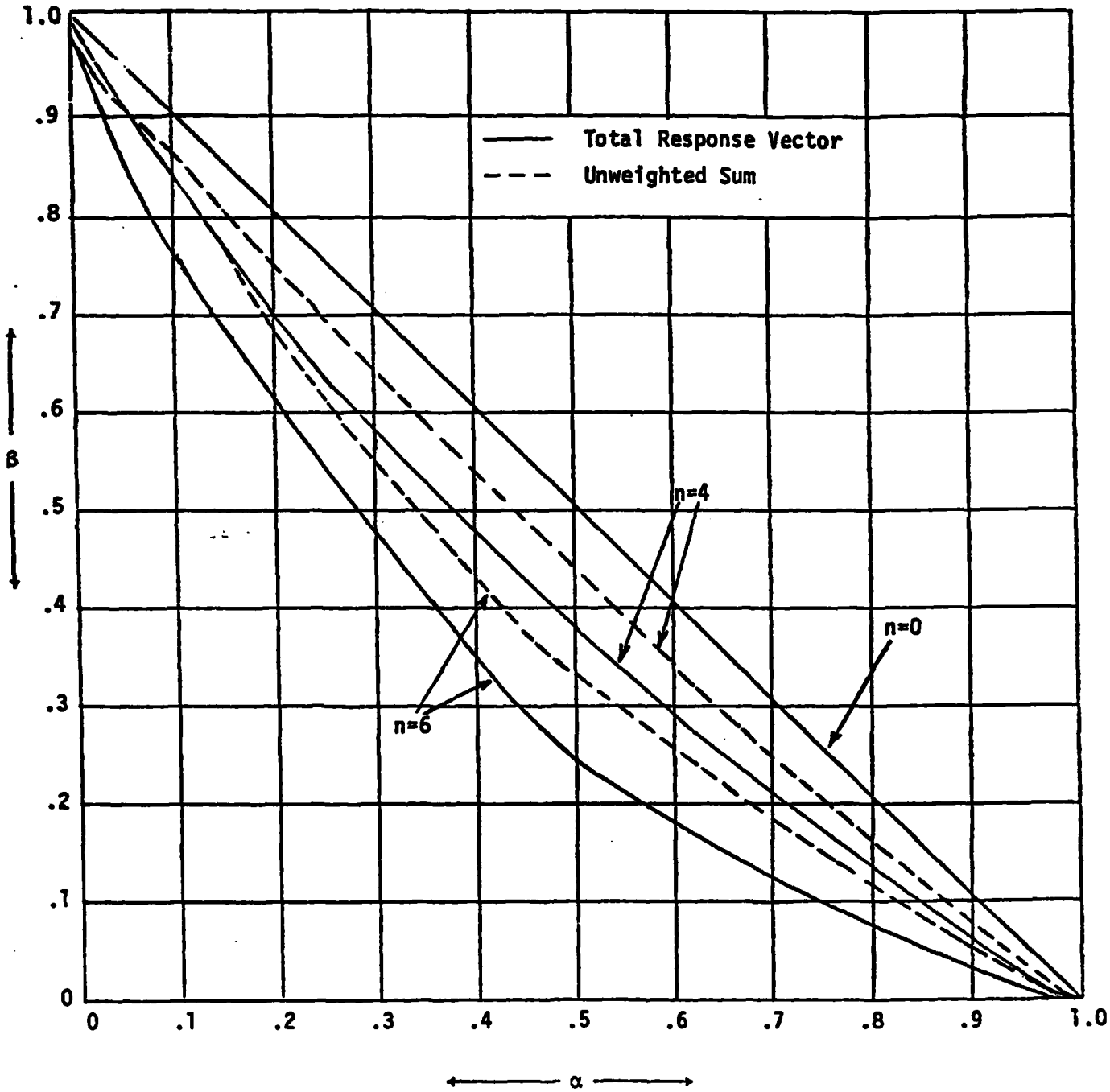
and classify a student with response vector x as a member of the high IQ group if and only if $L(x) \geq c$, otherwise classify as low IQ where $h(x)$ is the probability distribution of x in high IQ group and $\ell(x)$ is the probability distribution of x in low IQ group and c is a fixed constant equal to some value of $L(x)$. Now let α_c denote the probability of misclassifying a student from the high IQ group into the low IQ group when c is the cut-off point, that is

$$\alpha_c = \sum_{x:L(x)<c} h(x)$$

and likewise

$$\beta_c = \sum_{x:L(x)\geq c} \ell(x)$$

FIGURE 1


1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
0
β
Total Response Vector
Unweighted Sum
n=4
n=0
n=6
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1.0
α

is the probability of misclassifying an individual from low IQ group into high IQ group. For our four item case, there will be 17 decision rules, the i^{th} rule given by

$$\frac{h(x)}{\ell(x)} > c_i .$$

The c_i are determined in the following way. Take $h(x)/\ell(x)$ for all 16 response vectors, then realign them in descending order according to the values of the likelihood ratios. These values provide the cut-off points c_i which lead to 17 decision rules, the ith rule given by $h(x)/\ell(x)>c_i$. The first rule is classify as low-IQ, L, for every response vector; the second rule is classify in L unless the newly selected first response vector is observed; the third rule is classify in L unless one of the first two newly selected response vectors is observed, etc.; the seventeenth rule is classify as high-IQ, H, for every response vector. The risks in using each of these 17 rules are now computed by summing over the appropriate values of $h(x)$ and $\ell(x)$, which can be obtained from the new relisting. Obviously, the curve is always bounded by the straight line in the α, β plane connecting (1,0) and (0,1).

In Figure 1, we see the α,β curves for four and six items. If we give equal weight to α and β (the sample sizes for both IQ groups is 1491) the overall probability of misclassification ($\frac{1}{2}[\alpha+\beta]$) is approximately .44 when using the total response vector of four items, approximately .48 for the sum; approximately .39 when using the total response vector of six items, approximately .42 for the sum. One could say that administering six items and employing the sum is almost equivalent to administering four items and employing the total response vector.

The data base given in Table 1 provided an anchor for a number of investigators interested in the analysis of multivariate dichotomous data especially from a multi-dimensional contingency table or log-linear model viewpoint. Cox [2] gives a review of methods and models for analyzing multivariate binary data and lists the data in Table 1 as an example. Also, Martin and Bradley [7] applied a model based on a set of orthogonal polynomials to the data in Table 1 and Goodman [4] discusses the data in Table 1 as a base to examine methods for selecting models for contingency tables. We now present a procedure by Gokhale and Kullback [3] that is based on minimum discrimination information (MDI) estimation which they apply to the data in Table 1.

It may be instructive to look at this technique for the 2x2 contingency table.

Observed Values x(ij)

	j=1	j=2	
i=1	x(11)	x(12)	x(1.)
i=2	x(21)	x(22)	x(2.)
	x(.1)	x(.2)	x(..)=n

The dot is the label indicating summation over the index it replaces, that is x(1.) and x(2.) are the row marginals and x(.1) and x(.2) are the column marginals. Under the hypothesis of independence, the cell entries are estimated as a product of marginals, that is $x^*(ij) = x(i.)x(.j)/n$. Typically we then use chi-square with one degree of freedom to measure the divergence between x(ij) and $x^*(ij)$. Note the table of estimated values $x^*(ij)$ has the same marginals as the observed table x(ij). Now a common statistical measure of the association or interaction between the variables of a 2x2 contingency

table is the cross-product ratio or its logarithm which varies from $-\infty$ to $+\infty$ and is zero when there is no association.

For a 2x2x2 contingency table we have x(ijk) and an estimate under mutual independence is

$$x^*(ijk) = \frac{x(i..)x(.j.)x(..h)}{n^2} \quad .$$

One then asks whether these estimates or estimates under other hypotheses or models, e.g. first order interactions are not zero and the second order interaction is zero, provide good fits. An index to test goodness of fit is required and the χ^2 test statistic with an appropriate number of degrees of freedom emerges.

From information theory concepts, Gokhale and Kullback show that a log linear model results in which the log of the ratio of observed to fitted cells or fitted cells for two populations (as in our case) is a linear function of contingency cell parameters and that the measure of goodness of fit is chi-square. The cell estimates obtained in this fashion are called minimum discrimination information (MDI) estimates because the information index employed measures divergence between two distributions and this is to be minimized.

Gokhale and Kullback treat the data given in Table 1 as a five way 2x2x2x2x2 contingency table and wish to apply their technique to classification into one of the two multivariate dichotomous populations, that is, the same focus in the preceding discussion on classification. They denote the original observations in Table 1 by X(hijkℓ), where

Characteristic	Index	1	2
IQ	h	low IQ	high IQ
Item 1	i	disagree	agree
Item 2	j	disagree	agree
Item 3	k	disagree	agree
Item 4	ℓ	disagree	agree

As is typical, a study of main effects and interaction effects that impact on low IQ and high IQ discrimination is begun by examining the hypothesis that the IQ groupings are homogeneous over the 16 response vectors. The MDI estimate is of the form $X_a^*(hijk) = \frac{x(h....)x(.ijk\ell)}{n}$, the subscript a refers to this null hypothesis and the dot refers to summation over the index it replaces, and the test statistic

$$2I(x:x_a^*) = \frac{2\Sigma\Sigma\Sigma\Sigma x(hijk\ell)\log x(hijk\ell)\cdot n}{x(h....)x(\cdot ijk\ell)}$$

has a χ^2 distribution with 15 degrees of freedom. This is equivalent to the test for homogeneity in a 2x16 table. For the data in Table 1, $2I(x:x_a^*) = 68.369$ and so the hypothesis of homogeneity is rejected.

The next estimate x_b^* includes the marginal $x(hi...)$ and corresponds to the model that IQ is homogeneous over the response to the last three items, given the response to the first items, that is

$$x_b^*(hijk\ell) = \frac{x(hi\cdots)x(\cdot ijk\ell)}{x(\cdot i\cdots)}$$

TABLE 2

OBSERVED AND ESTIMATED SCIENCE ATTITUDE DATA

ijkℓ	Observed Low IQ $x(1ijk\ell)$	Estimated $x^*(1ijk\ell)$	Observed High IQ $x(2ijk\ell)$	Estimated $x^*(2ijk\ell)$
2222	62	74.589	122	109.414
2221	70	67.296	68	70.703
2212	31	31.329	33	32.671
2211	41	37.780	25	28.219
2122	283	266.570	329	345.429
2121	253	259.322	247	240.679
2112	200	193.625	172	178.376
2111	305	314.491	217	207.508
1222	14	12.156	20	31.844
1221	11	9.182	10	11.818
1212	11	9.659	11	12.341
1211	14	12.010	9	10.990
1122	31	33.623	56	53.375
1121	46	47.263	55	53.737
1112	37	47.450	64	53.550
1111	82	74.656	53	60.346
	1491		1491	

with

$$2I(x:x_b^*) = \frac{2\Sigma\Sigma\Sigma\Sigma\Sigma x(hijk\ell)\log\, x(hijk\ell)x(\cdot i\cdots)}{x(hi\cdots)x(\cdot ijk\ell)}$$

and from the data we get $2I(x:x_b^*) = 65.993$ with 14 degrees of freedom and therefore significance. Obviously more structure is required to get a good fit (that is, $2I(x:x^*)$ is not a significant χ^2 value). We thus continue in a sequential manner by examining additional hypotheses of main effects and interactions. The MDI estimates $x^*(hijk\ell)$ whose values are in Table 2 was selected because $2I(x:x^*) = 16.307$ with 11 degrees or freedom. This estimate x^* is symmetric with respect to the four items, and leads to the following parametric representation for the log-odds (low IQ/high IQ) over 16 response vectors. For example

$$\log \frac{[x^*(11111)]}{[x^*(21111)]} = \tau_1^h + \tau_{11}^{hi} + \tau_{11}^{hj} + \tau_{11}^{hk} + \tau_{11}^{h\ell}$$

that is, a linear regression for the log-odds in terms of an overall average τ_1^h and the main effects of each component of the response vector, namely τ_1^{hi}, τ_{11}^{hj}, τ_{11}^{hk}, $\tau_{11}^{h\ell}$. It is not surprising that the main effects alone lead to a good fit. In our previous discussion on using the total response vector, we computed all order interaction terms and their magnitudes were not very large suggesting they need not be included to obtain a good fit as measured by the χ^2 statistic.

From the data base in Table 1, the following values are obtained:

$$\tau_1^h = -0.3831 \qquad \tau_{11}^{hi} = -0.2030 \qquad \tau_{11}^{hj} = 0.1240$$

$$\tau_{11}^{hk} = 0.3411 \qquad \tau_{11}^{h\ell} = 0.3338$$

From the log-odds representation above we get

$$\frac{x^*(11111)}{x^*(21111)} = \exp(\tau_1^h)\exp(\tau_{11}^{hi})\exp(\tau_{11}^{hj})\exp(\tau_{11}^{hk})\exp(\tau_{11}^{h\ell})$$

or the odds ratio

$$\frac{x^*(11111)}{x^*(21111)} = (.682)(.816)(.1.132)(1.406)(1.396) = 1.237$$

leading to the probability that for a student who disagrees with all four items, there is a probability of $\frac{1.237}{2.237} = .55$ that he or she belongs to the low IQ category. From the actual observations, we get a probability of $\frac{82}{82+53} = .61$.

Since $x(1....) = x^*(1....) = 1491$ and $x(2....) = x^*(2....) = 1491$, we can assign a response vector $(ijk\ell)$ to population $h = 1$ (low IQ) or to population $h = 2$ (high IQ) depending on whether

$$\log \frac{x^*(1ijh\ell)}{x^*(2ijk\ell)} \geq 0$$

or

$$\log \frac{x^*(1ijh\ell)}{x^*(2ijk\ell)} < 0 .$$

The probability of error of misclassification from this assignment procedure is

$$\frac{1}{2}\left\{\sum_{(ijk\ell)\in 2} \frac{x^*(1ijk\ell)}{1491} + \sum_{(ijk\ell)\in 1} \frac{x^*(2ijk\ell)}{1491}\right\} .$$

This probability of misclassification error is 0.444. If we employ x instead of x^*, the probability is 0.441. Note that in our prior discussion of classification employing the total response vector and the likelihood ratio criterion, the overall probability of misclassification error was approximately 0.44, essentially, the same as it should be if the log-linear model employed is a good fit.

One important result of the use of the log-linear model to analyze attitudes toward science and IQ in high school seniors is the fact that odds ratios and consequently probabilities of events can be computed directly. For the classification problem where an assignment procedure is developed, the probabilities of misclassifiation can be computed but these are not as satisfying as say, the probability that a particular student is in the high IQ or low IQ category given a specific four item response on the questionnaire. It was this unconditional probability of an event that was desired in the 1950's but one settled for conditional probabilities within the classification framework and even there computer characteristics provided limitations.

Parole Outcome.

Let us now continue the use of the log-linear model on another data base where more than two category levels are permitted for some variables. In the early 1970's the US Parole Board was concerned with parole decision making. It would like parole to be granted a prisoner who would not violate parole conditions or be sentenced again at least for a two year period after release from prison. Since about thirty percent of those placed on parole are recidivists (return within two years), the Board was determined to learn what variables, if any, had an impact on recidivism and also the magnitude of the impact.

A study to accomplish this was initiated based on the records of approximately 2500 parolees (actually 2497) from federal prisons. For this group, 754 or about 30% were recidivists. A large number of variables were examined in connection with their association on recidivism and from these nine were selected, essentially by linear regression methods, to produce a score (linear sum) that would provide information on future failure or success in parole violation. Seven of these nine items and their levels are:

VARIABLES AND CATEGORY LEVELS: 2497 PAROLEES

	1	2	3
Prior Convictions:	none 338	one or two 609	three or more 1550
Prior Incarcerations:	none 779	one or two 726	three or more 992
Age at First Commitment:	18 or over 1503	under 18 994	
Commitment Offense:	auto theft 796	otherwise 1701	
Prior Parole:	no parole 1752	otherwise 745	
Drug History:	no hard drugs 1987	otherwise 510	
Release Plan:	with spouse 491	otherwise 2006	
Parole Outcome:	success 1743	failure 754	

Information on parole outcome is listed above with the seven items; the two items not appearing above are: employment record and educational level of prisoner.

In a study of this data by Solomon [9], four five-way contingency tables of variable factors believed to affect the outcome of parole were investigated. Of the four tables studied; one led to an estimated table relating the effects of four explanatory variables on parole outcome that was analyzed in detail. This estimate is based on a simple additive log-linear model, just as in the previous illustration on science attitude and IQ, but it accounts for a very high percentage of the total variation in parole outcome. The four explanatory variables in this case are: (1) number of prior convictions; notation H and index h; (ii) prior parole; notation J and index j; (iii) commitment offense; notation K and index k; (iv) release plan; notation M and index m - the parole outcome variable has notation N and index N.

The log odds representation that gives a good fit (93.2% explanation of the total variation) is

$$\log \frac{x^*(hjkm1)}{x^*(hjkm2)} = \tau^N + \tau_h^{NH} + \tau_j^{JN} + \tau_k^{KN} + \tau_m^{MN}$$

that is only the main effects of the four variables are required plus τ^N which is a general or unconditional measure of parole outcomes. In Table 2 we see observed and estimated parole outcomes employing the four way contingency table depicted in that listing. The estimated outcomes for each of the 24 patterns are close to the observed outcomes. This now permits the construction of Table 3 which shows for each pattern of item responses, the odds of parole success.

The probability of parole success for each pattern can then be easily obtained. For example, a prisoner with no prior convictions (level 1) and

TABLE 3

Number of Prior Convictions	Parole	Commitment Offense	Release Plan	Odds
1	1	2	1	15.727
1	1	1	1	10.108
1	2	2	1	9.840
2	1	2	1	7.425
1	1	2	2	7.279
1	2	1	1	6.324.
3	1	2	1	4.810
2	1	1	1	4.772
1	1	1	2	4.679
2	2	2	1	4.645
1	2	2	2	4.554
2	1	2	2	3.437
3	1	1	1	3.092
3	2	2	1	3.010
2	2	1	1	2.986
1	2	1	2	2.927
3	1	2	2	2.226
2	1	1	2	2.209
2	2	2	2	2.150
3	2	1	1	1.934
3	1	1	2	1.431
3	2	2	2	1.393
2	2	1	2	1.382
3	2	1	2	0.895

no prior parole (level 1) who is not in prison for auto theft (level 2) and has a release plan with spouse (level 1) has odds of parole success equal to 15.73 and a probability of parole success equal to $\frac{15.73}{16.73} = .94$; whereas a prisoner with a pattern (3,2,1,2) has odds of parole success equal to $\frac{.90}{1.90} = .47$. Recall that the observed odds of success, not conditioning on any explanatory variable is 2.31 leading to a probability of success equal to $\frac{2.31}{3.31} = .70$. Note that a prisoner with a (3,1,1,1) pattern, that is one with a large number of prior convictions but not too bad on the other variables has an odds success ratio = 3.092 and a probability of success equal to $\frac{3.092}{4.092} = .76$.

Thus those involved in parole decision making can get a quick idea of the probability of parole recidivism by responses to four items. This is even shorter than the nine item questionnaire obtained from regression analyses that leads to sums ranging between zero and eleven and moreover directly leads to a probability of an event of interest (parole success). In these ways the log-linear model far surpasses the regression or discriminant type analyses employed previously.

Towaway Accidents and Injuries.

Data was collected on approximately 3200 towaway accidents in the early 1970's, that is, autos in accidents that required towing to repair shops, see Solomon [10]. This accident data led to the following kind of multidimensional contingency table that included the following categorized variable.

(1) Severity of injury; categorized as either minor or none, or moderate or worse; (2) Auto occupant restraint system used; none, lap

and torso, and lap only. (3) General area of damage; undercarriage or top, right or left side, front, rear, unclassifiable. (4) Extent of impact; this was categorized into four components from very little to very heavy. The marginals for each variable follow.

THE MARGINALS FOR TOWAWAY ACCIDENT DATA

Restraint System Used		1st General Area of Damage	
None:	1777	Unclassifiable:	8
Lap Only:	566	Top or Bottom:	70
Lap and Torso:	866	Right or Left Side:	1193
		Front:	1719
		Rear:	219

Extent of 1st Impact		Injury	
One:	1234	Minor:	2832
Two:	1165	Worse:	377
Three or Four:	810		

The analysis then consisted of looking at a four-way (2x3x5x3) contingency table. The odds of a worse injury to a minor injury is what we wish to estimate. The log odds of a worse injury (2) to a minor injury (1) is given by

$$\log \frac{x^*_{2rdx}}{x^*_{1rdx}} = \lambda^I + \lambda^{IR}_r + \lambda^{ID}_d + \lambda^{IX}_x + \lambda^{IRD}_{rd} + \lambda^{IRX}_{rx} + \lambda^{IDX}_{dx} + \lambda^{IDRX}_{drx}$$

where I stands for injury, R for restraint, D for damage and X for impact. So we must first determine those effects which are non-zero, and then estimate the above log odds from the fitted model.

To determine the non-zero effects the following hypotheses were tested:

H_1: model includes only main effects (λ_r^{IR}, λ_d^{ID}, λ_x^{IX})

H_2: model includes main effects plus interaction between restraint used and general area of damage (λ_r^{IR}, λ_d^{ID}, λ_x^{IX}, λ_{rd}^{IRD})

H_3: model includes main effects plus interaction between restraint used and extent of first impact (λ_r^{IR}, λ_d^{ID}, λ_x^{IX}, λ_{rx}^{IRX})

H_4: model includes main effects plus interaction between general area of damage and extent of first impact (λ_r^{IR}, λ_d^{ID}, λ_x^{IX}, λ_{dx}^{IDX})

H_5: model includes main effects plus two interaction terms, the interaction between restraint used and extent of first impact and the interaction between area damaged and extent of first impact (λ_r^{IR}, λ_d^{ID}, λ_x^{IX}, λ_{rx}^{RX}, λ_{dx}^{DX})

H_6: model includes two main effects and one interaction, the effect due to general area of damage, the effect due to extent of first impact, and the interaction between them (λ_d^{ID}, λ_x^{IX}, λ_{dx}^{IDX}).

The following table indicates for each hypothesis the likelihood ratio statistic, the degrees of freedom, and the P-value associated with the likelihood ratio statistic.

	L.R.	D.F.	P-value
H_1:	2267	168	0
H_2:	364	56	0
H_3:	380	64	0
H_4:	64	56	.2203
H_5:	57	48	.1722
H_6:	811	60	0

Under H_1 we get a significant likelihood ratio value so that we reject H_1: Model includes only main effects. Therefore we must add an interaction term to the main effects. The three possible ways we could add an interaction are given by H_2, H_3, and H_4. All three hypotheses are quite an improvement over H_1 (as evidenced by the large decrease in the likelihood ratio value) but the best one by far is H_4, the addition of the interaction of general area of damage and extent of impact. Since we get a likelihood ratio of 64 with a P-value of .2203, we accept hypothesis H_4. This will be the model under which we estimate the effects. It accounts for 97% of the total variation in the original data.

Thus for the 3209 towaway accident cases in the study, the predicted odds ratio for major to minor injury is

$$e^{\lambda^{I} + \lambda_r^{IR} + \lambda_d^{ID} + \lambda_x^{IX} + \lambda_{dx}^{IDX}}$$

where the numerical estimates of the parameters lead to odds ratios and probabilities. The subscript values for λ relate to the categories given in the beginning of this section for I, R, D, X. Note that for this base, an interaction term, namely, damage and impact is indicated along with main effects to get a good fit.

Table 4 of of odds ratios and probabilities for minor injuries and worse injuries is now displayed. These are the operational results of the contingency table analysis on the towaway accident data.

TABLE 4

ODDS OF WORSE INJURY TO MINOR INJURY GIVEN THE TYPE OF RESTRAINT USED, FIRST GENERAL AREA OF DAMAGE, AND EXTENT OF IMPACT

FIRST GENERAL AREA OF DAMAGE	TYPE OF RESTRAINT	EXTENT OF IMPACT		
		1	2	3 OR 4
UNCLASSIFIABLE	NO RESTRAINT	0.4308	0.5874	1.2924
	LAP ONLY	0.2768	0.3775	0.8305
	LAP AND TORSO	0.1881	0.2565	0.5644
TOP OR BOTTOM	NO RESTRAINT	0.1958	0.7754	0.8285
	LAP ONLY	0.1258	0.4983	0.5323
	LAP AND TORSO	0.0855	0.3386	0.3618
SIDE	NO RESTRAINT	0.1023	0.1030	0.3175
	LAP ONLY	0.0657	0.0662	0.2040
	LAP AND TORSO	0.0446	0.0450	0.1386
FRONT	NO RESTRAINT	0.0785	0.1952	0.5643
	LAP ONLY	0.0504	0.1254	0.3626
	LAP AND TORSO	0.0343	0.0852	0.2464
REAR	NO RESTRAINT	0.0505	0.1119	0.0723
	LAP ONLY	0.0324	0.0719	0.0465
	LAP AND TORSO	0.0220	0.0488	0.0316

PROBABILITY OF A WORSE INJURY, GIVEN THE TYPE OF RESTRAINT USED, FIRST GENERAL AREA OF DAMAGE, AND EXTENT OF IMPACT

FIRST GENERAL AREA OF DAMAGE	TYPE OF RESTRAINT	EXTENT OF IMPACT		
		1	2	3 OR 4
UNCLASSIFIABLE	NO RESTRAINT	0.3011	0.3700	0.5637
	LAP ONLY	0.2168	0.2740	0.4537
	LAP AND TORSO	0.1583	0.2041	0.3607
TOP OR BOTTOM	NO RESTRAINT	0.1637	0.4367	0.4531
	LAP ONLY	0.1117	0.3325	0.3474
	LAP AND TORSO	0.0787	0.2529	0.2656
SIDE	NO RESTRAINT	0.0928	0.0934	0.2410
	LAP ONLY	0.0617	0.0621	0.1694
	LAP AND TORSO	0.0427	0.0430	0.1217
FRONT	NO RESTRAINT	0.0728	0.1633	0.3607
	LAP ONLY	0.0480	0.1114	0.2661
	LAP AND TORSO	0.0331	0.0785	0.1977
REAR	NO RESTRAINT	0.0481	0.1006	0.0674
	LAP ONLY	0.0314	0.0671	0.0444
	LAP AND TORSO	0.0215	0.0466	0.0306

Marine Reenlistment.

Another data base and the last to be discussed in this paper results from a study of Marine Corps reenlistment. Longitudinal data existed for about 10,000 Marines who enlisted in 1968. By 1972, information was then available on those who reenlisted and those who did not after service terms of 1, 2, 3, or 4 years. This presents an opportunity to examine the variables that might be associated with reenlistment and the impact of those variables on reenlistment. At first a large number of variables can be introduced that could possibly impact on reenlistment decisions, e.g. rank, military pay, number of dependents, length of enlistment, educational level, IQ, age, race, and area of the United States from which enlistment was accomplished. Some additional variables could be time before reenlistment decision at which rank is achieved, service in Vietnam or not, reenlistment bonus. Some of the more important variables turned out to be rank, length of enlistment, number of dependents, current primary job in the service, region of country from which Marine came, and IQ group. Some variables of a bit lesser importance turned out to be educational level, Vietnam service or not, and time at which rank is achieved. Such variables as race and age at enlistment turned out not to be important, but of course these could be associated with some of the important or somewhat important variables.

The overall reenlistment rate was .07. We would now like to examine profiles of Marines that might lead to much larger probabilities of reenlistment, and of course, those that would even give a smaller probability of reenlistment. In order to do this we once again rely on a on a log-linear model to obtain the log odds ratio of reenlistment to

nonreenlistment. When this is accomplished employing some of the more important variables we can offer some interesting results. For example we can write

$$\ln \frac{P_{1k}}{P_{2k}} = \beta^{T} + \beta_{k}^{TK}, \quad k = 1 \text{ for high rank, } k = 2 \text{ for low rank}$$

In this simple statement the log odds of reenlistment to non-reenlistment are seen to depend on β^{T}, the general mean for the log odds and β_{k}^{TK}, the association between rank and reenlistment decision.

To further illustrate what we now develop, consider another example. Assume that reenlistment is a function of two variables; length of enlistment, L, and the presence or absence of dependents, D. Then $P_{t\ell d}$ represents the probability that a specified reenlistment decision occurs given an individual's length of enlistment and dependency status. As before, the logarithm of the odds of reenlistment to not reenlisting can be written as

$$\ln \frac{P_{1\ell d}}{P_{2\ell d}} = \beta^{T} + \beta_{\ell}^{TL} + \beta_{d}^{TD} + \beta_{\ell d}^{TLD} .$$

Note that here we are allowing for the interaction of length of enlistment and dependency status.

We have already mentioned that the unconditional odds are .074 to one in favor of reenlistment. If we now condition on Marines who enlist for four years the odds of reenlistment increase from .074

TABLE 5

ODDS OF REENLISTMENT AND PROBABILITY OF REENLISTMENT

	Odds of Reenlistment	Probability of Reenlistment
Length of Enlistment		
Two Years	.041	.04
Three Years	.055	.05
Four Years	.182	.15
Race		
White	.047	.04
Non-white	.117	.10
Military Occupation		
Ground combat	.056	.05
Clerical and related	.073	.07
Other	.084	.08
General repair	.087	.08
Region		
East	.068	.06
North	.075	.07
South	.096	.09
West	.130	.12
Education		
High School or above	.061	.06
Less than High School	.090	.08
Combat		
In combat (Vietnam)	.076	.07
Not in combat	.104	.09

TABLE 6

ODDS OF REENLISTMENT AND PROBABILITY OF REENLISTMENT

A. Odds of Reenlistment by Length of Enlistment

	Length of Enlistment (in years)		
Rank	2	3	4
E1, E2	.014	.018	.079
E3	.019	.023	.095
E4	.037	.042	.266
E5 and above	.274	.258	.514
Dependents			
None	.030	.037	.072
One or more	.056	.080	.457
Current-Primary Job			
Same Current-Primary Job	.028	.054	.181
Different Current-Primary Job	.089	.087	.239

B. Probability of Reenlistment by Length of Enlistment

	Length of Enlistment (in years)		
Rank	2	3	4
E1, E2	.01	.02	.07
E3	.02	.02	.09
E4	.04	.04	.21
E5 and above	.22	.21	.34
Dependents			
None	.03	.04	.07
One or more	.05	.07	.32
Current-Primary Job			
Same Current-Primary Job	.03	.05	.15
Different Current-Primary Job	.08	.08	.19

to .182. Another result from this study shows that the odds of reenlistment for four years enlistees with one or more dependents equals .457 to one. If we omit the interaction term and thereby use only the main effects we would get a substantial underestimating, namely, .313 to one. If we continue in this way all kinds of odds or reenlistment can be developed as a function of predictor variables. Tables 5 and 6 indicate some of these odds as well as probability of reenlistment for various profiles. Many more reenlistment results from this large scale study can be obtained from Solomon, Haber and Ireland [11].

References

[1] Bahadur, R.R. (1961), 'A representation of the joint distribution of responses to n dichotomous items,' Studies in Item Analysis and Prediction, edited by Herbert Solomon, 158-168, Stanford University Press, Stanford, California

[2] Cox, D.R. (1972), 'The analysis of multivariate binary data,' Appl. Statist., 21(2), 113-120.

[3] Gokhale, D.V. and Kullback, (1978), The Information in Contingency Tables, Marcel Dekker, Inc., New York and Basel.

[4] Goodman, L.A. (1973), 'Guided and unguided methods for selecting models for a set of T multidimensional contingency tables,' J. Amer. Statist. Assoc., 68, 165-175.

[5] Imrey, P.B., Koch, G., and Stokes, M. (]98]), 'Categorical data analysis: some reflections on the log linear model and logistic regression. Part I: historical and methodological overview,' Int. Stat. Rev., 49, 265-283.

[6] Imrey, P.B., Koch, G., and Stokes, M. (1982), 'Categorical data analysis: some reflections on the log linear model and logistic regression. Part II: data analysis,' Int. Stat. Rev., 50, 35-63.

[7] Martin, D.C. and Bradley, R.A. (1972), 'Probability models, estimation, and classification for multivariate dichotomous populations,' Biometrics, 28, 203-221.

[8] Solomon, H. (1961), 'Classification procedures based on dichotomous response vectors,' Chapter 11, Studies in Item Analysis and Predictions, edited by H. Solomon, 177-186, Stanford University Press, Stanford, California.

[9] Solomon, H. (1976), 'Parole outcome, a multidimensional contingency table analysis,' Journal of Research in Crime and Delinquency, 107-126.

[10] Solomon, H. (1975), 'Passive restraint systems and accident outcomes,' Final Report to the U.S. Dept. of Transportation, DOT-HS-4-00974, April 30, 1975.

[11] Solomon, H., Haber, S. and Ireland, T. (1974), 'Manpower policy and the reenlistment rate,' Technical Report, TR-1201, June 10, 1974, the George Washington University.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 349	2. GOVT ACCESSION NO. A146620	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Log Linear Model Applications		5. TYPE OF REPORT & PERIOD COVERED TECHNICAL REPORT 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Herbert Solomon		8. CONTRACT OR GRANT NUMBER(s) N00014-76-C-0475
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-042-267
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics & Probability Program Code 411SP		12. REPORT DATE August 30, 1984 13. NUMBER OF PAGES 31
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Log linear models, Contingency tables, Multivariate data analysis

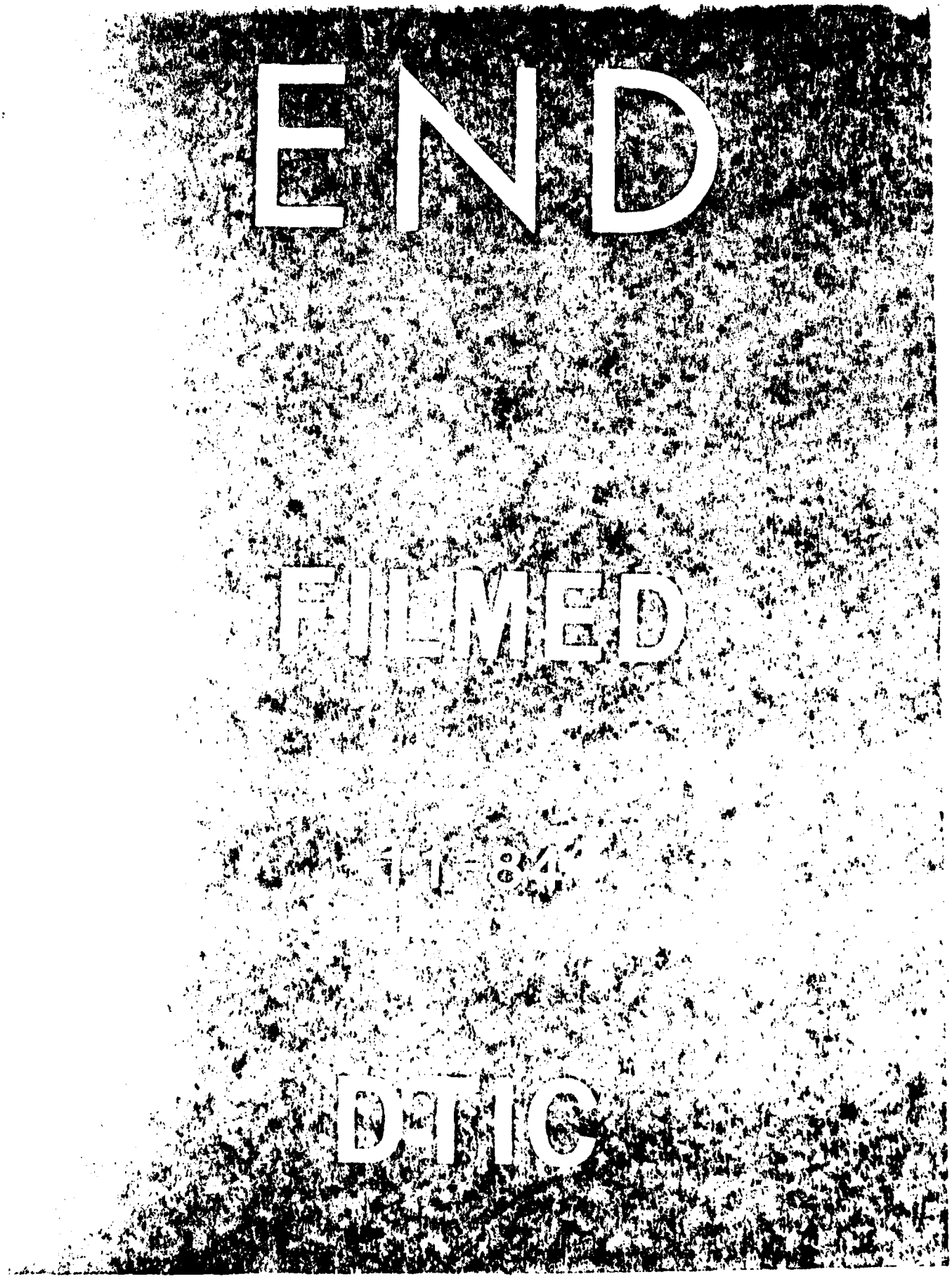
20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The log linear model is applied to several data bases. Among these are attitudes toward science in high school students, parole outcome, association of seat belts and injuries, and Marine re-enlistment association with Marine profile. Other multivariate techniques are discussed.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)


END
FILMED
DTIC