DOT-HS806-494
DOT-TSC-NHTSA-83-5

Analytical Methods in Multivariate Highway Safety Exposure Data Estimation

Peter Mengert
Edwin Roberts

Transportation Systems Center
Cambridge MA 02142

January 1984
Final Report

U.S. Department of Transportation
National Highway Traffic Safety Administration

Office of Research and Development
National Center for Statistics and Analysis
Washington DC 20590

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT-HS806-494		

4. Title and Subtitle
ANALYTICAL METHODS IN MULTIVARIATE HIGHWAY SAFETY EXPOSURE DATA ESTIMATION

5. Report Date
January 1984

6. Performing Organization Code
TSC/DTS-45

7. Author(s)
P. Mengert and E. Roberts

8. Performing Organization Report No.
DOT-TSC-NHTSA-83-5

9. Performing Organization Name and Address
U.S. Department of Transportation
Research and Special Programs Administration
Transportation Systems Center
Cambridge MA 02142

10. Work Unit No. (TRAIS)
HS470/R4418

11. Contract or Grant No.

12. Sponsoring Agency Name and Address
U.S. Department of Transportation
National Highway Traffic Safety Administration
Office of Research and Development
National Center for Statistics and Analysis
Washington DC 20590

13. Type of Report and Period Covered
Final Report
Oct. 1982-Nov. 1983

14. Sponsoring Agency Code
NRD-31

15. Supplementary Notes

16. Abstract

Three general analytical techniques which may be of use in extending, enhancing, and combining highway accident exposure data are discussed. The techniques are log-linear modelling, iterative proportional fitting and the expectation maximization (EM) method. A general discussion identifies a number of frequently encountered exposure data deficiences and indicates how one or more of the three analytical techniques may be of use in addressing each deficiency. The more mathematically-oriented sections provide a general introduction to each of the methods and discuss some of their properties of special interest in applications. A section illustrating applications to driving exposure data is included, together with computer program listings.

17. Key Words	18. Distribution Statement
Iterative Proportional Fitting, Log-Linear Modelling, Expectation Maximization, Accident Exposure Data (Highway), Categorical Data Analysis, Discrete Data Analysis, Iterations	DOCUMENT IS AVAILABLE TO THE PUBLIC THROUGH THE NATIONAL TECHNICAL INFORMATION SERVICE, SPRINGFIELD, VIRGINIA 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
UNCLASSIFIED	UNCLASSIFIED	134	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

PREFACE

This report presents the results of the first year of research on a task entitled "Analytical Alternatives to Multivariate Exposure Data Collection," conducted by the Transportation Systems Center (TSC) for the Mathematics Analysis Division of the National Center for Statistics and Analysis under the Office of Research and Development of the National Highway Traffic Safety Administration (NHTSA).

The authors thank Robert Arnold, the sponsor's technical monitor for useful suggestions and guidance.

METRIC CONVERSION FACTORS

Approximate Conversions to Metric Measures

Symbol	When You Know	Multiply by	To Find	Symbol
		LENGTH		
in	inches	*2.5	centimeters	cm
ft	feet	30	centimeters	cm
yd	yards	0.9	meters	m
mi	miles	1.6	kilometers	km
		AREA		
in^2	square inches	6.5	square centimeters	cm^2
ft^2	square feet	0.09	square meters	m^2
yd^2	square yards	0.8	square meters	m^2
mi^2	square miles	2.6	square kilometers	km^2
	acres	0.4	hectares	ha
		MASS (weight)		
oz	ounces	28	grams	g
lb	pounds	0.45	kilograms	kg
	short tons (2000 lb)	0.9	tonnes	t
		VOLUME		
tsp	teaspoons	5	milliliters	ml
Tbsp	tablespoons	15	milliliters	ml
fl oz	fluid ounces	30	milliliters	ml
c	cups	0.24	liters	l
pt	pints	0.47	liters	l
qt	quarts	0.95	liters	l
gal	gallons	3.8	liters	l
ft^3	cubic feet	0.03	cubic meters	m^3
yd^3	cubic yards	0.76	cubic meters	m^3
		TEMPERATURE (exact)		
°F	Fahrenheit temperature	5/9 (after subtracting 32)	Celsius temperature	°C

*1 in = 2.54 (exactly). For other exact conversions and more detailed tables, see NBS Misc. Publ. 286, Units of Weights and Measures, Price $2.25, SD Catalog No. C13.10:286.

Approximate Conversions from Metric Measures

Symbol	When You Know	Multiply by	To Find	Symbol
		LENGTH		
mm	millimeters	0.04	inches	in
cm	centimeters	0.4	inches	in
m	meters	3.3	feet	ft
m	meters	1.1	yards	yd
km	kilometers	0.6	miles	mi
		AREA		
cm^2	square centimeters	0.16	square inches	in^2
m^2	square meters	1.2	square yards	yd^2
km^2	square kilometers	0.4	square miles	mi^2
ha	hectares (10,000 m^2)	2.5	acres	
		MASS (weight)		
g	grams	0.035	ounces	oz
kg	kilograms	2.2	pounds	lb
t	tonnes (1000 kg)	1.1	short tons	
		VOLUME		
ml	milliliters	0.03	fluid ounces	fl oz
l	liters	2.1	pints	pt
l	liters	1.06	quarts	qt
l	liters	0.26	gallons	gal
m^3	cubic meters	35	cubic feet	ft^3
m^3	cubic meters	1.3	cubic yards	yd^3
		TEMPERATURE (exact)		
°C	Celsius temperature	9/5 (then add 32)	Fahrenheit temperature	°F

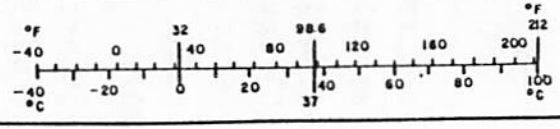

TABLE OF CONTENTS

TABLE OF CONTENTS (CONT.)

LIST OF ILLUSTRATIONS

LIST OF TABLES

LIST OF TABLES (CONT.)

LIST OF ABBREVIATIONS

DF	degrees of freedom
E-M (EM)	Expectation - Maximization
FHWA	Federal Highway Administration
IPF	Iterative Proportional Fitting
LLM	Log-Linear Modelling
NHTSA	National Highway Traffic Safety Administration
NPTS	Nationwide Personal Transportation Survey
TSC	Transportation Systems Center
VMT	Vehicle Miles Travelled

1. INTRODUCTION

This report presents the results of the first year of research on a task entitled "Analytical Alternatives to Multivariate Exposure Data Collection," conducted by the Transportation Systems Center (TSC) for the National Center for Statistics and Analysis of the National Highway Traffic Safety Administration (NHTSA).

Multivariate exposure data are used in conjunction with accident data to produce estimates of accident rates for various subsets of highway use, determined by different combinations of driver, vehicle, and road characteristics, and environmental conditions. This type of detailed information is particularly needed in crash avoidance and accident causation analysis and can be used to confirm serious safety problem areas and to aid in the evaluation of countermeasures. In contrast to the collection of multivariate accident data, which is routinely done by reporting requirements and follow-up studies, the collection of useful multivariate exposure data usually requires the implementation of expensive, large-scale surveys. The purpose of this task is to develop ways to use mathematical analysis to obviate the need for primary collection of multivariate exposure data.

The specific goal of this research effort is to develop methods to estimate needed "large" dimensional multivariate exposure tables from data sources of smaller dimension and/or more limited context. In general, existing multivariate exposure data are limited in geographic coverage or time period, are generated from a small sample and are otherwise limited in detail or accuracy. Often, exposure data of significantly different types or non-exposure related data (e.g., census, registration, accident data) are all that is available.

The key methodological problem is to expand and integrate the disparate data sources to develop "best" estimates of multi-dimensional exposure tables in ways which maintain relationships among data elements. Because there are

so many possible combinations of variables that could conceivably be of interest, the research effort has focused on the development of statistically sound and demonstrably valid methods rather than on the generation of a few specific tables.

Section 2 "An Overview of Exposure Data Problems and Analytical Remedies," provides an overview of the type of exposure data set that is commonly used in traffic safety analysis, and of the common deficiencies encountered in existing exposure data. A preliminary description is given of the major analytical methods for treating such data problems, and there is discussion of which methods may be most appropriate for particular problems.

Section 3 "Discussion of Analytical Techniques," is the heart of the report, containing a detailed mathematical description of the major analytical techniques, a summary of their properties derived from the literature, and a description of new results, relevant to exposure data analysis, which were discovered in the course of this research. Section 3 also contains a discussion of important statistical considerations in the context of exposure data applications.

Section 4 contains some simple, preliminary examples of applications of several techniques to exposure data.

2. AN OVERVIEW OF EXPOSURE DATA PROBLEMS AND ANALYTICAL REMEDIES

This section describes some of the common deficiencies encountered in existing highway safety exposure data (i.e., VMT),* and a brief overview of how certain analytical techniques might be used to improve the situation. A more extensive treatment of these techniques, their properties, and their potential for mitigating exposure data problems is given in Section 3. Some preliminary applications of these techniques to exposure data are shown in Section 4. The major issues to be addressed in Section 2 are:

- How existing exposure data fall short of fulfilling multivariate exposure data needs.

- Which analytical methods have potential for mitigating some of the shortcomings in existing exposure data.

- Which techniques are relevant to what data problems.

Table 1 displays a matrix whose row headings are a list of prototypical problems which plague existing exposure data. The column headings refer to mathematical techniques which seem to have the highest potential for mitigating these problems. An X is entered in a cell where a particular technique is believed to have application to a particular problem.

Section 2.1 provides a background description of the type of data sets under consideration, and a brief overview of the principal relevant analytical techniques. Sections 2.2 through 2.8 contain brief discussions of each problem area and how the various techniques may be employed to mitigate that problem.

*It has been assumed throughout this report that the exposure data measure of interest is vehicle miles travelled (VMT).

TABLE 1. DATA PROBLEMS AND REMEDIAL TECHNIQUES

Common Defects/Deficiencies in Multivariate Exposure Data \ Possible Remedial Techniques	Log-Linear Modelling	IPF	E-M Algorithm
1. Incomplete Classification of Exposure Data		X	X
2. Small Samples - Insufficient Cell Counts	X		X
3. Presence of Zero Cells	X		X
4. Missing Levels (Categories)			X
5. Classification of Levels Inconsistent Across Samples			X
6. Incorrect Data	X		
7. Wrong Location or Timeframe		X	

2.1 COUNTED DATA SETS AND ANALYTIC METHODS

The data under consideration are known generically as "counted" data. They are arranged in arrays, each cell of which contains the number of individuals from an underlying population (or sample) classified by certain characteristics. (Accident data generally conform to this type of data. Exposure data also conform to the classification aspects of counted data, but the cell entries (VMT) more closely resemble continuous data. More is said in Section 3.2.4 regarding adjustments needed to apply the theory of counted data to exposure data.) The characteristics are described by means of variables and levels of variables. For example, the population might be the licensed drivers in a State. One variable might be gender, with levels male and female. Another variable might be age, with levels 0-15, 15-35, 35-55, over 55. A typical arrangement for such a two-variable set would be a 2x4 matrix whose rows correspond to the two levels of gender, whose columns correspond to the four levels of age. In general, a $k_1 x k_2 x \ldots x k_n$ array is an n-dimensional array whose ith dimension has k_i levels. In the following section, the terms "core" and "margin" will be used frequently. By "core" we always mean a data set (real or dummy) having the full set of variables (and usually the full set of levels) of interest. "Margins" are defined as data sets that can be derived from a core by summing over all the levels of one variable or all the levels of several variables. For example, the margins of a matrix are the 2 one-dimensional arrays corresponding to the row sums and the column sums, respectively. Note that we include under the term "margin," arrays which correspond to a subset of variables and their levels of a particular core, but have cell counts observed independently of that core. Thus, the cell counts do not equal the cell counts of the corresponding margin summed from the core. When it is necessary to distinguish such margins, they are referred to as "exogenous" margins.

The following mathematical techniques are briefly discussed below, in order to provide background for the remainder of this Section: Log-linear Modelling (LLM), Iterative Proportional Fitting (IPF) and Dempster's Expectation-Maximization Algorithm (E-M). A thorough discussion of these techniques, including their definitions and reference to prior literature, is presented in Section 3.

Log-linear modelling is the most natural general method for fitting a model to counted data: It can be used for data smoothing, including estimating the value of zero cells, and for detecting outliers. (See Section 3.2 for a complete description).

IPF is a method of fitting a core to a set of (compatible) exogenous margins. For example, in the two-dimensional case, given an nxm matrix M with positive cell counts, an n-vector A and an m-vector B with positive entries such that the sum of the entries of A equals the sum of the entries of B, the procedure consists of alternately scaling: the rows of M so that the new sums equal the corresponding entries of A, the columns of M so that the column sums equal the corresponding entries of B. The procedure usually converges when all the input arrays are positive. A complete description of IPF, including properties of convergence, uniqueness of solutions and a mathematical characterization of the qualitative relationship of the solution to the input core is given in Section 3.3.

The E-M algorithm is a very general procedure for obtaining maximum likelihood estimates of cell counts, due to Hartley* and Dempster.**(See Section 3.4). Its application to our problems is a systematic way of pooling different observations of the same data (in the form of cores and margins) by a process that includes alternately estimating a log-linear model based on the current expected values of the cell counts and updating the expected value of the cell counts based on the current parameters of the model. A complete technical description of the method and the result of our investigation of such issues as convergence and uniqueness is given in Section 3.4.

The sections below discuss the relevance of these techniques to the various problem areas.

*See Reference 7.
**See Reference 4.

2.2 INCOMPLETE CLASSIFICATION OF EXPOSURE DATA

In this case the situation is that exposure for the population of interest needs to be (or is desired to be) classified, jointly, by a certain set K of variables. However, each of the several existing samples is classified according to some proper subset of K, so that there is no fully classified sample. Let the number of variables by which a sample S is classified be called the dimension of S, and suppose that the desired dimension of a fully classified sample is k. In this case, there are basically two analytical approaches to enhancing the data in a manner aimed at approximating a fully classified sample. Both involve piecing together the existing data sets (each of a dimension less than k) to form a data set of dimension k.

In the first case:

- Assume that unobserved interactions among variables are zero. Iterative Proportional Fitting (IPF) can then be used to estimate the completely classified sample using a k-dimensional (dummy) core of all 1's, after adjusting the available data sets for use as (compatible) margins. The Expectation-Maximization (E-M) algorithm can also be used in this case, if an unsaturated* log-linear model (of order not greater than the dimension of the largest sample) is used. As the theory described in Section 3.2.2 will reveal, the former (IPF) approach actually results in the maximum likelihood estimation of a log-linear model for any core having the given margins. As discussed in Section 3.5, however, the latter approach (EM) is probably the preferred method in this case.

In the second case:

- Utilize information on the unobserved interactions from other sources (e.g., a fully classified sample from another time or place). In this case, the only apparently practical method is to fit the margins to a

*See Section 3.2 for definition.

core, containing interactions approximating the unobserved interactions, via IPF. An interesting implication of the theory expounded in Section 3.3.3 is that it is only these interactions (in the core) that have any effect on the outcome of IPF, i.e., the given core could be replaced in the replaced in the computation by a core with identical unobserved interactions and an arbitrary set of interactions corresponding to the observed (in the margin) interactions, and the outcome array of IPF will be the same. Thus, the resulting data set will contain the interactions present in the margins together with higher level interactions derived from the surrogate core. (See Section 3.3.3 for a more detailed discussion.) The E-M algorithm could be applied formally to obtain a unique solution in this case for the saturated or unsaturated log-linear model, but it would not be clear what the outcome of EM would mean. Hence, EM is not a desired method for this case.

Various complications can be present in all these cases due to basic deficiencies in the available samples themselves. The major anticipated problems appear to be:

- The core is unstable due to insufficient observations (cell counts).
- The core or marginal samples contain too many zero cells.
- Certain levels (categories) of some variables were not observed in the core or in some of the marginal samples.
- Classification of levels for some variables is not consistent across samples.
- Incorrect data is present in one or more samples.
- Completely classified data are available but from the wrong location or time frame.

Each of these complications is a data deficiency in its own right, which analytical methods may or may not be able to mitigate. Each is described in more detail in the following sections.

2.3 SMALL SAMPLES - INSUFFICIENT CELL COUNTS

The situation is that a (fully classified) multivariate sample of exposure (or accident) data consists of a small number of observations so that, at least for desired levels of disaggregation, cell counts have unacceptably large standard errors. There appear to be two possible approaches to mitigating this problem:

(a) Fit an unsaturated log-linear model to the data set. While some information (probably of dubious value due to the small sample) on higher order interactions will be sacrificed, this process should reduce the standard error in the cell counts.

(b) If marginal samples (incompletely classified observations) of reasonably large size, from the same population, are available, the E-M algorithm can be used to systematically combine the samples to fit log-linear model.

Note the similarity between the situation discussed in this section and the case of piecing together incompletely classified samples in the presence of a fully classified core, described in section 2.2. The difference lies in the following facts. In the former case, the focus of interest is on the margins, which are the only direct observations of the population of interest. The core plays the surrogate role of supplying subjectively adequate estimates of interaction information missing from the marginal samples. In this case there need be no statistical evidence that the core is sampled from the same population as the margins. In the latter case, the focus of interest is the core (which contains an insufficient number of observations). The objective here is to find a valid way of pooling marginal data to reduce the standard error of the core cells. In this case, it is imperative to have evidence that the core and marginal samples belong to the same population.

2.4. PRESENCE OF ZERO CELLS

Where cells in a multivariate table are not structural zeros (e.g., a cell representing the number of motorcycles with weight = 4000 lbs.), a sample may still have a zero cell count because of the combined effect of a limited number of observations and a relatively small count for that cell in the entire population. In this event it can be useful to have a technique for estimating the correct relative count for that cell from the given population. The fitting of an unsaturated log-linear model should give such an estimate for zero cell counts. If marginal information is present, it may be useful to fit an unsaturated log-linear model using the E-M algorithm.

Additional complications that can be caused by zeros in the core or marginal samples are: failure of IPF to converge, and non-uniqueness of the outcome of the E-M algorithm (dependence of the outcome on the starting array). This indicates that, where IPF is used, considerable smoothing of core and marginal tables by means of unsaturated log-linear models will probably be required. (In the case of the E-M algorithm, the analogous adjustment is to avoid the saturated model when too many zeros are present in the input tables).

2.5. LEVELS OF VARIABLE(S) MISSING

This situation can best be illustrated by means of a (fictitious) example: suppose that VMT is to be classified jointly by driver age and sex, vehicle type, vehicle age and day vs. night. Suppose further that there is available a special study which observed all these variables, but only for drivers under 24 years of age. The possible age categories for drivers over 24 are the "missing levels" for the variable called "driver age." This is an incomplete data situation, one for which the E-M algorithm was designed. Carrying the example a bit further, suppose that, for the population of interest, VMT has been estimated by driver age and day/night jointly, and also by driver age and vehicle age jointly (for all levels of variables). Theoretically, the E-M algorithm can be used to obtain maximum likelihood estimates for the fully classified data set, using a saturated (or

unsaturated) log-linear model for those cells where age is less than 24 and an unsaturated model for those cells where age is over 24. The extent to which missing levels can be successfully mitigated by E-M or other analytical techniques is not clear. After preliminary analysis as to whether such problems might arise significantly in practice, the situation should be carefully investigated, if warranted.

2.6 CLASSIFICATION OF LEVELS (I.E., BOUNDARIES OF CATEGORIES) FOR VARIABLE(S) NOT CONSISTENT ACROSS SAMPLES

Resorting again to an example, suppose we have a fully classified multivariate core of exposure data, (e.g., VMT), and suppose there are also partially classified, independently observed margins. Suppose further that one of the variables is driver age, and that the levels (in the margin) for each sample are: 15-24, 25-40, 40-55, 55-65, over 65. Suppose that the classification of driver age in the core is 15-24, 25-55, and over 55. This is a case where the classification of levels are incompatible but commensurate, in fact the levels (categories) in the core are aggregates (set-theoretic union) of the levels (categories) in the margins. The case where levels are incommensurate as well as incompatible may be a more difficult problem. This is illustrated by taking the marginal levels for driver age as given in the example above, and supposing that the core has as levels: 15-30, 30-45, 45-60, and over 60. The E-M algorithm is theoretically capable of addressing the problem of commensurate incompatible levels, and we feel it could be extended to cover the incommensurate case as well. As with the problem described in section 2.5, it is not obvious how successful analytical techniques might be in mitigating this problem, and the extent to which this problem can be tolerated or corrected for.

2.7 OUTLIERS

The final type of problem is the presence of errors in the data. One of the established uses of log-linear models for categorical data tables (the form of all data of interest in this study) is to detect outliers. This is done by fitting log-linear models to the data set of interest and checking

the relative goodness of fit of each cell. The usefulness of this check and the determination of effective ways of dealing with detected outliers is an area that may warrant some investigation.

2.8 EXPOSURE DATA FROM WRONG TIME OR PLACE

The situation is that partially classified (marginal) data sets exist for the population of interest, and a fully classified data set (core) exists for all the variables of interest, but the core is not from the same population as the margins because the core is from an earlier time frame, or a different geographic locale. This is a special occurence of the second case described in Section 2.2, where IPF is recommended if the core (i.e., the core interaction observed in the margins) is thought to be a reasonable surrogate for the time or place of interests. Confidence in the outcome of the result can be enhanced by testing (if data are available) for stability of the unobserved interaction over time or place.

3. DISCUSSION OF ANALYTICAL TECHNIQUES

3.1 INTRODUCTION

In this chapter three general analytical procedures for dealing with categorical or cross-classified data are described and some of their properties discussed. The three procedures are:

1. Log linear modeling
2. Iterative proporational fitting (IPF)
3. The expectation-maximization (EM) "algorithm"

The three procedures are related. Log linear modelling will be discussed first since its concepts and techniques play a role in the other two. Each of the three procedures involves similar concepts and will use some common notation. They will all be discussed in terms of similar simple hypothetical examples.

The hypothetical examples will refer to a completely classified matrix Y_{ijk}, of count data. The matrix Y_{ijk} is called count data since it takes only non-negative integral values, it is called completely classified since it gives a count corresponding to specific values of each of the "variables" indicated by i j and k (which are the entire set of variables in this three dimensional example). Thus, variable 1 has been indexed by i, variable 2 by j and variable 3 by k. Variable 1 could, for example, be driver age, variable 2 driver sex, and variable 3 could be vehicle type. Y_{ijk} could be a count of vehicles classified by these three variables. Variable 1 will be assumed to have the levels 1, 2, ..., I. This means that i can take on the values 1 through I. Variable 2 will have J levels (1 through J) and variable 3 will have K levels. Less than completely classified data sets (or marginal data sets) will also be of interest. For example, R_{ij} might be a table of counts classified by only the first two variables (for each observation the third variable is assumed unknown). Similarly S_k will denote a table of counts classified by only the third variable. The margins of Y_{ijk} are also examples of less than completely classified data. For example, Y_{ij+} will

represent the margin of Y_{ijk} formed by summing over the third variable, k.

Thus $Y_{ij+}=\sum_{k=1}^{K} Y_{ijk}$. Other margins are similarly represented e.g.

$$Y_{++k} = \sum_{i=1}^{I} \sum_{j=1}^{J} Y_{ijk} \text{ and } Y_{i+k} = \sum_{j=1}^{J} Y_{ijk}$$

Although the log linear modelling, IPF and EM procedures will be described in this report in terms of a hypothetical 3 dimensional completely classified matrix with one and two dimensional margins of interest, the techniques are applicable to more general arrays with some unusual exceptions which will be noted. It will be convenient to keep the notation simple to the greatest possible extent but it should be borne in mind that the Y matrix could have any number of subscripts, e.g., Y_{ijklm} and any number of these can be summed over to provide marginal matrices of interest, e.g., Y_{i++l+}.

3.2 LOG LINEAR MODELS

Log linear models have received far more extensive treatment in literature than the other analytical techniques reviewed in this report (IPF and EM). Reference to the books by Bishop and Goodman (reference 1 and 2) is sufficient to indicate the impressive literature on the subject. The fundamentals are however essentially simple and are reviewed briefly below.

3.2.1 Definition

Let Y_{ijk} denote a matrix of counted data i.e. each entry is an integer. Further let $Z_{ijk}=E(Y_{ijk})$ *and assume that Y_{ijk} is distributed as a Poisson random variable with mean Z_{ijk} or as a multinominal random variable with $Y_{+++}=Z_{+++}=N$ and $p_{ijk}=Z_{ijk}/N$. Then a log linear model for Y_{ijk} corresponds to a multiplicative form for Z_{ijk}. The saturated log linear model is of the form

$$Z_{ijk} = \alpha_i\beta_j\gamma_k\delta_{ij}\,\varepsilon_{jk}\phi_{ki}\,\mathcal{H}_{ijk} \qquad 3.1$$

(This is no restriction on Z_{ijk}.) The second order unsaturated model sets $\mathcal{H}_{ijk}=1$:

$$Z_{ijk} = \alpha_i\beta_j\gamma_k\,\delta_{ij}\,\varepsilon_{ik}\,\phi_{ki}.$$

*The notation "E( )" denotes the expected value of the quantity in the parentheses.

The (homogeneous) first order unsaturated model states that in addition δ_{ij}, $\varepsilon_{jk}, \phi_{ki}$ all equal one so: $Z_{ijk} = \alpha_i \beta_j \gamma_k$. These are examples of hierarchical log linear models, a term which will now be explained.

Consider the saturated model for Z_{ijk} and take logarithms of both sides of the defining equation:

$$\log Z_{ijk} = U_o + U_{1(i)} + U_{2(j)} + U_{3(k)} + U_{12(ij)} + U_{23(jk)} + U_{13(ik)} + U_{123(ijk)}. \quad 3.2$$

The U's in the above equation are indeterminate since there are a total of

$$1+I+J+K+IJ+JK+IK+IJK$$

of them while there are a total of IJK of the quantities $\log Z_{ijk}$. The ambiguity is resolved by requiring that any U summed over any of its subscripts yields zero, e.g.

$$U_{1(+)} = 0, U_{12(i+)} = 0, U_{123(+jk)} = 0, \text{ etc.} \quad 3.3$$

Then the U's are completely determined by equation 3.2 and the zero summation conditions. In what follows when any matrix is expressed in the form (3.2) (with some of the U's possibly missing or set equal to zero) it is to be understood that the zero summation conditions hold. With this notation $U_{123(ijk)}$ measures the third order interaction (in Z_{ijk}) (or the "three factor effect"). Similarly, for example, $U_{23(jk)}$ represents a specific two factor effect and $U_{1(i)}$ a specific one factor effect.*

Hierarchical (log linear) models are characterized by the condition that whenever an effect is zero, all higher order effects involving all the variables in the zero effect must also be zero.

---

*In this report the terms "interaction" and "effect" will be synomymous as will "two factor effect" and "second order interaction." The term "two factor effect" with respect to the model expressed by equation 3.2 will refer to $U_{12(ij)}$ or $U_{23(jk)}$ or $U_{13(ik)}$ where for example, $U_{12(ij)}$ represents a set of terms indexed by i and j.

For example,

$$\log Z_{ijk} = U_o+U_{1(i)}+U_{2(j)}+U_{3(k)}+U_{12(ij)}+U_{23(jk)}$$

is a hierarchical model. The key missing (zero) effect is $U_{13(ik)}$. Since it is missing, $U_{123(ijk)}$ must be zero also. Consider also:

$$\log Z_{ijk} = U_o+U_{1(i)}+U_{2(j)}+U_{3(k)}+U_{12(ij)} \quad , \qquad 3.4$$

this is also a hierarchical model since no low order effect is missing with a higher order effect involving its variables present. Consideration of log linear models in this report will be limited to hierarchical log linear models. One example of a nonhierarchical model is given for illustration:

$$\log Z_{ijk} = U_o+U_{1(i)}+U_{2(j)}+U_{13(jk)}$$

3.2.2 Maximum Likelihood Estimation Using IPF

Under the assumption of multinomially distributed Y_{ijk}, a maximum likelihood estimate of Z_{ijk} is obtained using Iterative Proportional Fitting or (IPF). This process will be described in more generality in Section 3.3 but its application to log linear models is described here.

To fit a given log linear model to Y_{ijk} i.e. to find a maximum likelihood estimate of Z_{ijk} given Y_{ijk}, the procedure will be given. X_{ijk} will denote the estimate to be derived. Let the model* to be fit be given by

$$\log Z_{ijk} = U_o+U_{1(i)}+U_{2(j)}+U_{3(k)}+U_{12(ij)} \quad .$$

The procedure is as follows:

1. Form the margins of Y_{ijk} corresponding to the highest order effect involving each variable. So in this case form Y_{++k} (corresponding to $U_{3(k)}$) and Y_{ij+} (corresponding to $U_{12(ij)}$).

2. Initialize the matrix which will be iteratively scaled to give the final estimate of X_{ijk} :

*A brief discussion will be given subsequently of the process of choosing a model.

Set $X_{ijk}^{(o)} = 1$

3. Scale to each margin successively and iterate the process:

$$\text{Set } X_{ijk}^{(2n+1)} = (Y_{ij+}/X_{ij+}^{(2n)}) X_{ijk}^{(2n)}$$

and then set

$$X_{ijk}^{(2n+2)} = (Y_{++k}/X_{++k}^{(2n+1)}) X_{ijk}^{(2n+1)}$$

for n = 0, 1, 2, 3, ...

4. When convergence is reached X_{ijk} is found:

$$X_{ijk} = \lim_{n \to \infty} X_{ijk}^{(n)}$$

(Convergence is assured, see Section 3.3).

This process will be recognized as a special case of IPF. In the notation to be used in Section 3.3 on IPF, X_{ijk} represents the result of applying IPF to a core, in this case $M_{ijk} = 1$, with margins $R_{ij} = Y_{ij+}$ and $S_k = Y_{++k}$. Incidentally, if structural zeroes are required in the model, they are inserted into the matrix M_{ijk}.

3.2.3 Statistics For Goodness of Fit

The fit of a log linear model is assessed by either the X^2 or the G^2 statistic using the appropriate number of degrees of freedom. Before describing how to compute X^2 and G^2 the computation of the number of degrees of freedom is given.

Each term in a log linear model has a specific number of degrees of freedom associated with it: e.g. $U_{12(ij)}$ has (I-1)(J-1) degrees of freedom while $U_{123(ijk)}$ has (I-1) (J-1) (K-1) degrees of freedom and $U_{3(k)}$ has K-1 degrees of freedom. The parameter U_o (e.g. in equation 3.2) has one degree of freedom. The degrees of freedom for the model equals the sum of the degrees of freedom of its terms. For example, consider the degrees of freedom (DF) for the model in equation 3.4, i.e.

$$\log Z_{ijk} = U_o + U_{1(i)} + U_{2(j)} + U_{3(k)} + U_{12(ij)} \qquad 3.5$$

For this model:

$$DF=1+(I-1) + (J-1) + (K-1) + (I-1)(J-1) = K+IJ-1 \qquad 3.6$$

The statistic X^2 or G^2 has a chi square distribution with degrees of freedom equal to IJK minus the number of degrees of freedom in the model i.e.,

$$IJK-(K+IJ-1) = IJK-K-IJ+1$$

in the preceding example.

The definitions of G^2 and X^2 are as follows:

$$X^2 = \sum_{ijk} (X_{ijk} - Y_{ijk})^2/X_{ijk} \qquad 3.7$$

$$G^2 = 2 \sum_{ijk} Y_{ijk} \log (X_{ijk}/Y_{ijk}) \qquad 3.8$$

The sums are over all cells in the original data matrix, Y_{ijk}, except, in the case of G^2, for cells where $Y_{ijk} = 0$, in which case the contribution to G^2 for those cells is zero.

If two models are under consideration and the second contains all the effects that the first does plus one or more added effects, then a decision to choose between them may be based on the difference in X^2 (or G^2) divided by the difference in degrees of freedom. The more complex model is chosen if the difference in X^2 (G^2) is statistically significant (too large to be reasonably due to chance) for the given difference in degrees of freedom. The difference in X^2 (G^2) will have a chi square distribution with degrees of freedom equal to the difference of degrees of freedom of the models [under the null hypothesis that the extra effects in the more complex model are not needed (i.e. have the value zero)].

The process of log linear modelling consists of these steps:

1. Select a set of candidate log linear models.

2. Fit the selected models to the data.

3. Drop from consideration any model which is:
   a. a special case (i.e. a restriction) of a more complex model which has a significantly smaller X^2 (G^2);
   b. an extension of a simpler model which does not have a significantly larger X^2 (G^2).

4. Consider additional models which contain more effects as suggested by the effects present in the retained models.

In summary, the key tools for estimating and evaluating log linear models are:

1. Margin building.

2. IPF.

3. Computation of G^2 and/or X^2.

3.2.4 Discussion of Problems Encountered in Applying Log Linear Models to Continuous Data

In general, when exposure is to be calculated, the sum of VMT over each cell is the variable of interest. This is not a "counted" variable as required in the development of the standard statistical techiques for log linear models. The cumulative VMT in a cell is not a count variable both because it is a cell sum of a (practically) continuous quantity and because sample weighting factors are applied (in the case of NPTS data at least). Even if trips were analyzed instead of VMT, the individual trips would not be independent (in the case of the NPTS data) and hence the assumptions needed in the development of standard statistical techniques for log linear models would not be satisfied. In general, standard discrete multivariate statistical modelling techniques are developed under assumptions not satisfied by VMT data.

In this section, the log linear modelling problem is considered from a point of view which recognizes that the cell measures involved in some cases (in particular, VMT) are not counts but instead sums of random numbers of terms each of which is positive and for practical purposes continuously distributed (the individual terms may be weighted trip lengths for example).

The problem is then how to model VMT cross-classified by several variables. Clearly log linear models provide a useful mathematical form for describing cumulative VMT classified, for example, by driver, vehicle, roadway and environment classes. This is because log linear models are useful for describing non-negative quantities in which a multiplicative model for the joint effects of factors is of interest. In short, the log linear model structure is just as useful whether the quantity in each cell is a cumulative sum (of continuous terms) or is instead a cumulative count.

Other aspects of the log linear modelling process as developed for count data are problematic however. In particular, the maximum likelihood method for fitting the model must be examined and the statistical criteria for model adequacy must also be examined.

With regard to the method of fitting the model, two questions need to be addressed:

1. What is the form of the likelihood in the case of continuous cell sums instead of cell counts, and to what extent is maximizing it approximately accomplished by the classical discrete method.

2. Is the classical fitting method useful in providing satisfactory fits in the more general continuous case?

In Appendix A it is shown that the classical fitting method (i.e. applying IPF to margins as done in the log linear modelling process) leads to log linear models fit to the data according to a closeness of fit criterion which is reasonable and this is independent of the statistical properties of the data. This is a situation analogous to the application of linear regression in cases where the statistical properties assumed for the residuals do not hold (even though a linear relationship between the variables is postulated).

The question of whether the fitted model remains a maximum likelihood estimate is examined and it is concluded that a maximum liklihood estimate is obtained if these assumptions hold:

1. The cell sums of VMT are normally distributed (this will be so if each cell sum is over many records).

2. The ratio of variance to mean is constant across cells.

Assumption 1 is probably valid to the necessary degree. Assumption 2 probably is not valid to a substantial degree. However, if the two assumptions were valid, the chi square statistics obtained in the course of the standard log linear modelling process could be modified by a single factor (the same for any model based on the initial data) easily calculated from the original data. The factor can be calculated in any case by the formula given in Appendix A. The validity of such a calculation is discussed in Appendix A.

Since log linear models are appropriate in the continuous case (e.g. when cell values are VMT) and since the classical method of estimating log linear models (developed for the discrete case) leads to models of closest fit by a reasonable criterion, the chief weakness of applying the classical or standard method in the general case is the problematic nature of the statistical fit criterion (chi square values in the classical case) used to decide what degree of model complexity is justified by the data. It should be noted that except where the assumptions needed to make the classical approach give true maximum likelihood estimates in the general case hold, it appears to be a nearly impossible job to develop methods for producing true maximum likelihood estimates in the general situation. This point is brought out in Appendix A.

It is concluded that log linear models will remain of great interest in the general case. It is also likely that they may effectively be estimated using the classical method of applying IPF to selected margins. However, the statistical criterion used in model development based on chi square statistics will need to be modified. The modified version (as described in Appendix A) may not be entirely satisfactory in all cases. Other methods for assessing statistical stability such as the jackknife and infinitesimal jackknife* techniques or other methods based on split samples may be of use. If a general robust method for assessing statistical stability is developed, it can be used in conjunction with the modified chi square (i.e. chi square multiplied by the correction factor) to gain experience on the validity of the latter (which should be much easier to use than general robust methods).

*See Reference 8.

3.3 ITERATIVE PROPORTIONAL FITTING (IPF)

The second analytical technique to be dealt with in this report is Iterative Proportional Fitting or IPF. It can be a useful technique when marginal data is available relating to the context of interest and complete data is available relating to a somewhat different context. The analytical technique and some of its properties will be discussed in this section.

3.3.1 General Discussion

In IPF, a completely classified matrix (of any number of dimensions) is initialized to some "core" matrix which determines the high order interactions. In the course of the computation, the matrix is successively scaled to match in turn each of a set of marginal matrices and the scaling process is repeated (i.e., iterated) until it converges. The resulting matrix matches each of the marginal matrices and as noted has its higher order interactions determined by the initial core matrix.

In the next section, three equivalent formulations of IPF will be given, utilizing a representative hypothetical example similar to the one considered in Section 3.2

Various properties of the IPF technique will be discussed in relation to one or another of the equivalent formulations.

Before proceeding to give the three equivalent formulations, the procedure will be defined in the usual way by describing the computational process involved. This will be the same as the third of the equivalent formulations to be given in the next section. The technique is illustrated in terms of the representative example:

> Suppose M_{ijk} is given as the (initial) core matrix and R_{ij} and S_k are given as marginal matrices. The result of the IPF process applied to these data will be denoted by the matrix X_{ijk}. Let $X^{(o)}_{ijk}$ denote the value at the initial step. By definition of the IPF process $X^{(o)}_{ijk} = M_{ijk}$. Then the current $X^{(n)}_{ijk}$ matrix is scaled to each of the margins in turn and the scaling process is repeated until convergence is obtained

(below there will be a discussion of cases where convergence is not obtained). Suppose that X_{ij+} is to match R_{ij} and X_{++k} is to match S_k. This is achieved through repetitively scaling the X_{ijk} matrix until agreement with both incompletely classified or marginal matrices (i.e. R_{ij} and S_k) is reached (to a prespecified accuracy).

An algebraic formulation of the process is as follows:

1. Set $X_{ijk}^{(o)} = M_{ijk}$ (for all values of i, j, k)

2. For each n in turn (starting with n=0) update X_{ijk} as follows:

   Set

   $$X_{ijk}^{(2n+1)} = X_{ijk}^{(2n)} \left(R_{ij}/X_{ij+}^{(2n)}\right)$$

   then set

   $$X_{ijk}^{(2n+2)} = X_{ijk}^{(2n+1)} \left(S_k/X_{++k}^{(2n+1)}\right)$$

   (repeat for n=0, 1, 2, ..., etc.)

3. $X_{ijk}^{(n)}$ converges to X_{ijk}:

   $$\lim_{n\to\infty} X_{ijk}^{(n)} = X_{ijk}$$

   (Cases under which convergence may not occur are discussed below.)

3.3.2 Three Equivalent Formulations for IPF

IPF as described in the previous section is equivalent to two other formulations given here. In the statement of the theorem below the previous formulation is repeated for completness. The problem statements define the same problem in the sense that each calls for finding a matrix* which has certain properties, satifies certain conditions, and/or is constructed in a certain way. Each of the three formulations leads to the same matrix X_{ijk} (or else none of the problems can be solved due to infeasible conditions).

The three formulations are as follows:

Formulation 1: Find X_{ijk} of the form $X_{ijk} = \alpha_{ij} \beta_k M_{ijk}$ such that

$$X_{ij+} = R_{ij}$$

$$X_{++k} = S_k$$

Formulation 2:** Find $\underline{X}$ which minimizes

$$F(\underline{x}) = \sum_{ijk} (X_{ijk} \log_e (X_{ijk}/M_{ijk}) - X_{ijk})$$

subject to the constraints:

$$X_{ijk} = R_{ij}$$

$$X_{++k} = S_k$$

$$X_{ijk} > 0$$

Formulation 3: Find

$$X_{ijk} = \lim_{n \to \infty} X_{ijk}^{(n)}$$

where

$$X_{ijk}^{(o)} = M_{ijk}$$

*The formulations are given in terms of the standard example but all statements carry over to more general dimensionalities except where noted.

**See, e.g., Reference 3.

and

$$X_{ijk}^{(2n+1)} = X_{ijk}^{(2n)} R_{ij}/X_{ij+}^{(2n)}, \quad X_{ijk}^{(2n+2)} = X_{ijk}^{(2n+1)} S_k/X_{++k}^{(2n+1)} \text{ for } n = 0,1,2, \ldots$$

(Formulation 3 is merely the formulation given in the previous section.)

3.3.3 Discussion on the Consequences of Formulation 1 of IPF

Formulation 1 states that $X_{ijk} = M_{ijk}\alpha_{ij}\beta_k$ where α_{ij} and β_k are uniquely determined by the margin conditions ($X_{ij+} = R_{ij}$, $X_{++k} = S_k$) (The uniqueness is a property of IPF addressed in Section 3.3.4).

Some consequences of this formulation of IPF related to log linear models will now be noted. In order to state these consequences precisely, define the following saturated log linear models:

Let

$$\log M_{ijk} = U_o^M + U_{1(1)}^M + U_{2(j)}^M + U_{3(k)}^M + U_{12(ij)}^M + U_{23(jk)}^M + U_{13(ik)}^M + U_{123(ijk)}^M \quad 3.9$$

$$\log R_{ij} = U_o^R + U_{1(i)}^R + U_{2(j)}^R + U_{12(ij)}^R \quad 3.10$$

$$\log S_k = U_o^S + U_{3(k)}^S \quad 3.11$$

$$\log X_{ijk} = U_o^X + U_{1(i)}^X + U_{2(j)}^X + U_{3(k)}^X + U_{12(ij)}^X + U_{23(jk)}^X + U_{13(ik)}^X + U_{123(ijk)}^X \quad 3.12$$

The interactions in M_{ijk} which do not correspond to any interactions in R_{ij} or S_k will be called "margin-absent" interactions (or effects).* Thus, the last three terms in equation 3.9 represent margin-absent effects. The effects in M_{ijk} which correspond to effects represented in one or more of the data margins will be called "margin-present" effects. Thus the first five terms in equation 3.9 represent margin-present effects. In general, the margin-absent effects are higher order effects based on factors whose lower order effects are margin-present effects. (Thus the margin-absent effects

*The term "effect" is synonymous with "interaction" (see Section 3.2.1).

can be spoken of loosely as higher order effects and margin-present effects as lower order effects.) The first characteristic of IPF to be observed is:

1. The result, X_{ijk}, of performing IPF with core data M_{ijk} and margin data R_{ij} and S_k does not depend on the margin-present effects in M_{ijk}. That is, if the margin-present effects in M_{ijk} are changed in any way, the result, X_{ijk}, is unchanged.

2. The second observation is:
   The margin-absent effects in X_{ijk} are equal to the margin-absent effects in M_{ijk}. (Roughly speaking, the higher order interactions in the initial core are preserved intact.)

   To these observations, a third obvious one may be added:

3. The margins of X_{ijk} are equal to the corresponding marginal data matrices (R_{ij} and S_k respectively).

Properties 1, 2, and 3 are characteristic of IPF. Any procedure characterized by properties 2 and 3 is equivalent to IPF. The importance of the properties lies in the fact that cores for IPF are characterized completely by their margin-absent effects. The margin absent effects of a matrix can be isolated and compared to those of another matrix to determine if they are equivalent for use in IPF. Each matrix M_{ijk} can be converted to a reduced matrix M^*_{ijk} where $\log M^*_{ijk}$ equals the sum of the terms in 3.9 corresponding to margin-absent effects. Then M^*_{ijk} does not contain (non zero values for) the margin-present effects and so represents an equivalent core to M_{ijk} which can be compared to the reduced matrix of another core matrix.

The remainder of this section is devoted to brief outline of the proof of the assertions regarding the characteristics of IPF given earlier in this section.

The second stated property (margin-absent effects in the core are preserved) will be derived first:

Since $X_{ijk} = \alpha_{ij}\beta_k M_{ijk}$.

we have:

$$\log X_{ijk} = U^X_o + U^X_{1(i)} + U^X_{2(j)} + U^X_{3(k)} + U^X_{12(ij)} + U^X_{23(jk)}$$

$$+ U^X_{13(ik)} + U^X_{123(ijk)} = \log \alpha_{ij} + \log \beta_k + U^M_o$$

$$+ U^M_{1(i)} + U^M_{2(j)} + U^M_{3(k)} + U^M_{12(ij)}$$

$$+ U^M_{23(jk)} + U^M_{13(ik)} + U^M_{123(ijk)}$$

and

$$\log \alpha_{ij} = U^\alpha_o + U^\alpha_{1(i)} + U^\alpha_{2(j)} + U^\alpha_{12(ij)}$$

$$\log \beta_k = U^\beta_o + U^\beta_{3(k)}$$

Then effects of various types must be equal individually* so that e.g. $U^X_o = U^M_o + U^\alpha_o + U^\beta_o$ and in particular $U^X_{123(ijk)} = U^M_{123(ijk)}$

$$U^X_{23(jk)} = U^M_{23(jk)},\ U^X_{13(ik)} = U^M_{13(ik)}$$

which state the equality of the margin-absent effects in X_{ijk} and M_{ijk} (property 2).

The first property (that margin-present effects in the core are completely inconsequential) is derived as follows:

Let M'_{ijk} be like M_{ijk} in its margin-absent effects but different in its margin-present effects.

Then

$$\log M'_{ijk} = U^{M'}_o + U^{M'}_{1(i)} + U^{M'}_{2(j)} + U^{M'}_{3(k)} + U^{M'}_{12(ij)}$$

$$+U^{M'}_{23(jk)} + U^{M'}_{13(ik)} + U^{M'}_{123(ijk)} \qquad 3.13$$

*Since the representation of a positive matrix as a saturated log linear model is unique.

where the last three terms of $\log M'_{ijk}$ equal the corresponding terms of $\log M_{ijk}$ but the other terms are different. Then

$$\log M'_{ijk} = \log M_{ijk} + a_{ij} + b_k$$

where a_{ij} and b_k are arbitrary.

Let $X'_{ijk} = \alpha'_{ij}\ \beta'_k\ M'_{ijk}$

and $X'_{ij+} = R_{ij}$ and $X'_{++k} = S_k$

Then since*

$$M'_{ijk} = M_{ijk}\exp(a_{ij})\exp(b_k)$$

we have

$$X'_{ijk} = \alpha'_{ij}\ \exp\ (a_{ij})\ \beta'_k\ \exp\ (b_k)\ M_{ijk}$$

By the uniqueness of IPF (and since X'_{ijk} is now expressed in the form $\alpha_{ij}\beta_k M_{ijk}$) we must have

$$X'_{ijk} = X_{ijk}\ \text{(QED)}$$

That properties 2 and 3 characterize IPF may be shown as follows:
Let X'_{ijk} be a matrix characterized by properties 2 and 3 i.e. such that $X'_{ij+} = R_j$ and $X'_{++k} = S_k$ and such that the margin-absent effects in X'_{ijk} are the same as those in M_{ijk}. Then $X'_{ijk} = \delta_{ij}\ \varepsilon_k\ M_{ijk}$

and by the uniqueness property of IPF, $X'_{ijk} = X_{ijk}$.

---

*Notation: $\exp\ (x) = e^x$

3.3.4 Discussion of the Consequences of Formulation 2 of IPF

Formulation 2 may be called the mathematical programming formulation. It constitutes an optimization problem with a strictly convex objective function and linear constraints. Let

$$f(u) = u \log_e u - u \qquad 3.14$$

Then the mathematical programming formulation is:

$$\underset{\underline{x}}{\text{Min}} \sum_{rst} M_{rst} f(X_{rst}/M_{rst}) \qquad (= F(\underline{x})) \qquad 3.15$$

subject to the constraints

$$\begin{aligned} X_{ij+} &= R_{ij} \\ X_{++k} &= S_k \\ X_{ijk} &\geq 0 \end{aligned} \qquad 3.16$$

(This assumes $M_{ijk} > 0$; some modifications in the following observations are needed if any of the M_{ijk} are zero. IPF is not recommended in such a case unless it is a structural zero.)

It is easy to show that F is strictly convex in the X_{rst}'s (since $d^2f(u)/du^2 < 0$).

A mathematical programming problem with linear constraints and a strictly convex objective function either has a unique solution or else is infeasible (i.e. the constraints cannot be satisfied) and has no solution. This shows that there is a unique solution to IPF (convergence of the IPF algorithm will be dealt with in Section 3.3.5) so long as the margins are feasible. Unfortunately it is possible to have infeasible margins either because they are incompatible (e.g. if $R_{++} \neq S_+$) or essentially infeasible, i.e. compatible but still infeasible. The latter condition is dealt with in Section 3.3.5 and in in Appendix B. It is not expected to be a common problem. In any case, there can be no more than one solution to any of the three equivalent formulations of IPF.

The math programming formulation thus provides a useful assurance of uniqueness of the solution as well as an indication of the importance (and sufficiency) of feasibiltiy.

Perhaps more important is the characterization provided of the meaning of the result of the IPF procedure. Figure 1 shows a plot of f(u) vs. u. It has a minimum at u = 1. Thus minimizing the objective function F ($\underline{X}$), consists in drivng X_{ijk} as close as possible to M_{ijk}, subject to the marginal constraints. Other objective functions can be proposed with the same constraints, but the resulting formulations are not necessarily equivalent to IPF. Examples of different objective functions with the same constraints (i.e. those in 3.16) leading to procedures different from IPF include an example in reference 3 and a special case of EM given in Section 3.4.3.

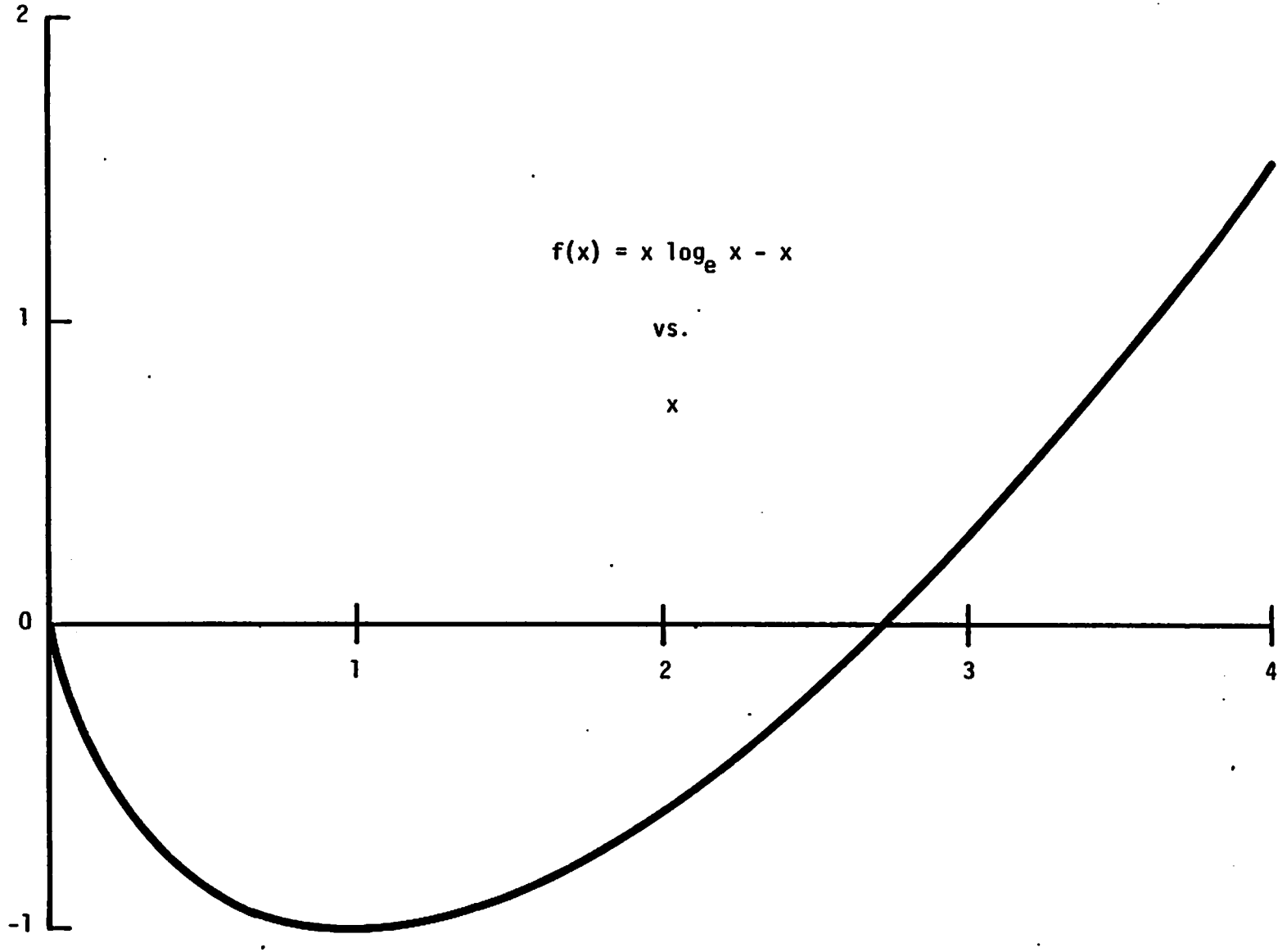

FIGURE 1. GRAPH OF A FUNCTION

3.3.5 Discussion of the Consequences of Formulation 3 of IPF

Formulation 3 is simply the specification of the calculation procedure as given in Section 3.3.1. A convergence proof is given in Reference 1 (Bishop et al). The proof given there is for the case of IPF used to fit log linear models, but appears to be extendable to the case of general IPF if the assumption of the existence of a feasible solution is included (Reference to other proofs may be found in Reference 1 and Reference 6).

Appendix B discusses the general problem of criteria for the existence of feasible solutions. An example of compatible but infeasible margins is given there.

In practice convergence will take place quickly (in the IPF procedure) if the margins are feasible. Convergence is characterized by the condition that $X_{ijk}^{(n)}$, for sufficiently large n stops changing to any appreciable extent i.e.

$$|X_{ijk}^{(n+d)} - X_{ijk}^{(n)}|$$

is very small for d = 1, 2, 3, ...*

If the margins are infeasible, $X_{ijk}^{(n)}$ will cycle instead of converging.** The length of the cycle is equal to the number of margins being scaled to. In the example where two margins are being scaled to this means that $|X_{ijk}^{(n+2)} - X_{ijk}^{(n)}|$ is very small but $|X_{ijk}^{(n+1)} - X_{ijk}^{(n)}|$ is not small. The matrix X_{ijk} comes back to the same place each time it is scaled to a particular margin but wanders away when scaled to another margin.

The condition of stable cycling will indicate infeasible margins, but this condition can be broken down into two subcategories.

1. Incompatible but otherwise feasible ("simply incompatible") margins.

2. Essentially infeasible margins.

*A convergence criterion based on comparing the margins of $X_{ijk}^{(n)}$ with their target values can also be used, and has some advantages.

**Cycling sets in gradually, i.e., is established asymptotically.

In the first case, "simply incompatible margins," the problem is simply that the margins do not agree along common (sub-) margins. In this case if the disagreement is not too large a satisfactory compromise can be reached by choosing $X_{ijk}^{(n)}$ after scaling to the most significant margin or taking the average of $X_{ijk}^{(n)}$ over a cycle.

In the case of infeasible margins, the marginal matrices may even be compatible but a serious inconsistency exists such that no <u>non-negative</u> matrix has the given marginal matrices as its margins. It appears that then the $X_{ijk}^{(n)}$ matrix will develop zeroes where no zeroes are found in the margins. No satisfactory IPF solution can be found in this case.

See Appendix B for more discussion of the margin conditions.

Some general guidelines can be established for detecting and dealing with anomolies in the computational procedure.

1. In the case of feasible, compatible margins, convergence will be rapid.

2. If the margins are feasible but not compatible, stable cycling of the algorithm will be observed with no zeros present in the $X_{ijk}^{(n)}$. In this case:

   a. If the swings in the cycle are not large, an average or best solution (of the cycling outcome matrices) may be an adequate solution.

   b. If the swings in the cycle are large, then the margins may be too incompatible to represent the same population* (recall that the <u>core</u> need not represent the same population).

3. If the margins are essentially infeasible, cycling will occur, with zeros appearing in certain cells of $X_{ijk}^{(n)}$. In this case, IPF is not an appropriate treatment of the given data.

*Marginal discrepancies in scale are likely to be less troublesome than marginal descrepancies in distribution. Discrepancies in scale can be eliminated altogether by scaling all margins to the same grand total before commencing IPF.

3.4 THE EM METHOD

The Expectation Maximization Method or "EM Algorithm" was expounded by Dempster, Laird and Rubin in 1976 (Reference 4). The EM method is a general means of finding maximum likelihood estimates of model parameters in the face of incomplete data. Incomplete data is data to be used in estimating a model which actually models some (additional) data or information which is not available. For example, if some of the observations in the data lack (i.e. are missing) information on some of the variables considered by the model, this would be a case of incomplete data. A typical example would involve log linear models to be estimated from samples which contained some observations which were fully classified and some which were not classified by one or more of the variables of interest in the model.

3.4.1 General

The expectation maximization method (for estimating models using incomplete data) assumes that all the data are governed by the same underlying distribution (as embodied in the modelling assumptions). An iterative process is undertaken to estimate (certain of the) parameters of the distribution (i.e. the model parameters). The technique uses initial estimates of model parameters to estimate the complete data by an expectation process (the E step) and then uses the estimated complete data to find a maximum likelihood estimate of the model parameters (the M step). Then the model parameters which result from the M step are used again to estimate complete data in the E step and the steps are repeated alternately to convergence. There is a certain intuitive appeal to the procedure, but more importantly Dempster et al showed that the process as they describe it leads to actual maximum likelihood estimates pertinent to the actual data and model at hand.

All the applications to be considered are expected to be representable by exponential family type distributions (certainly this is the case for log linear models for count data). Dempster et al have developed a particularly elegant formulation for the EM method in the case where the data is assumed to

arise from an exponential family distribution. An exponential family distribution* is one for which the density function has the form

$$f(x|\phi) = b(x) \exp(\phi t(x)^T)/a(\phi) \qquad 3.17$$

Here x represents the random variable describing the (hypothetical) complete data, (x, ϕ and t(x) are multivariate i.e. each may be a vector),** ϕ represents the parameters of the distribution of x for which it is desired to find a maximum likelihood estimate; t(x) is a sufficient statistic for ϕ given x, $a(\phi)$ and b(x) are arbitrary functions subject to the requirements of probability distributions.

The EM algorithm states that a maximum likelihood estimate of ϕ can be obtained by iterating the following two steps:

1. The E step:

$$t^{(p)} = E(t(x)|y, \phi^{(p)}) \qquad 3.18$$

Here $t^{(p)}$ is referred to as an "estimate" of the complete data. In the above expression y denotes the actual incomplete data (y is known, x if known would determine y by a many to one mapping of x). $\phi^{(p)}$ represents the estimate of ϕ as it stands at the p^{th} iteration. Under favorable conditions as p gets larger $\phi^{(p)}$ will approach a maximum likelihood estimate of ϕ.

2. The M step:

Determine a maximum likelihood estimate of ϕ based on t_p; this will be $\phi^{(p+1)}$. Then go back to the E step with $\phi^{(p+1)}$ in place of $\phi^{(p)}$

*Familiarily with the exponential family is not required in the following discussion.

**Dempster's notation is used here. Elsewhere in this report vectors and matrices are indicated by subscripts or underlines. Note that $t(x)^T$ here denotes the transpose of the vector t(x).

3.4.2 Application to Log Linear Models for Incompletely Classified Data

A translation of this formal specification in the particular case of some completely classified and incompletely classified data matrices of count data will give a better idea of the procedure.

Let the data be given in the form of matrices of count data, M_{ijk}, R_{ij} and S_k. Then the M_{ijk} data is said to represent a completely classified sample while R_{ij} and S_k are incompletely classified in that the observations that make them up (i.e. the individuals counted in the matrices) are not classified by all three variables (only by the first two variables in the case of R_{ij} and only by the third variable in the case of S_k). Thus R_{ij} and S_k in this sense represent incomplete data. Let R'_{ijk} represent the (hypothetical) complete data for R_{ij} and S'_{ijk} represent the complete data for S_k. Then $R_{ij} = R'_{ij+}$ and $S_k = S'_{++k}$.

Now M_{ijk}, R'_{ijk}, S'_{ijk} represent the complete data for determination of the model parameters. As in the usual incomplete data problem the complete data set is not actually known since in place of R'_{ijk}, R_{ij} is given and in place of S'_{ijk}, S_k is given. Since in the general statement of the technique x was used to represent the complete data, M_{ijk}, R'_{ijk}, S'_{ijk} will take the place of x in the procedure. Similarly the incomplete data will be M_{ijk}, R_{ij}, S_k (recall that the incomplete data was denoted by y in the general exposition).

Let*

$$T_{ijk} = M_{ijk} + R'_{ijk} + S'_{ijk} \qquad 3.19$$

Then T_{ijk} is a sufficient statistic for**

$$\phi_{ijk} = E(T_{ijk}) = E\,(M_{ijk} + R'_{ijk} + S'_{ijk}) \qquad 3.20$$

*Note that T_{ijk} represents the total number of observations in cell ijk.

**The ϕ_{ijk} are the model parameters in this problem. This is discussed in more detail below.

Under the usual assumptions of log linear modelling T_{ijk} is Poisson with mean ϕ_{ijk} or multinomial (with $P_{ijk} = \phi_{ijk}/T_{+++}$).

Since M_{ijk}, R'_{ijk}, S'_{ijk} are all from the same population it will be assumed that M_{ijk} is Poisson with mean $\alpha\phi_{ijk}$, R'_{ijk} with mean $\beta\phi_{ijk}$ and similarly S'_{ijk} with mean $\gamma\phi_{ijk}$ (α, β, γ are not known.)

Under these assumptions*

$$E(R'_{ijk}|\phi_{ijk}, R_{ij}) = (\phi_{ijk}/\phi_{ij+}) R_{ij} \qquad 3.21$$

This represents part of the E step since R_{ij} is part of the incomplete data, ϕ_{ijk} denotes the model parameters, and R'_{ijk} will provide an additive part of the sufficient statistic.
The other equation needed is:

$$E(S'_{ijk}|\phi_{ijk}, S_k) = (\phi_{ijk}/\phi_{++k}) S_k \qquad 3.22$$

This then leads to:

$$E(T_{ijk}|\phi_{ijk}, M_{ijk}, R_{ij}, S_k) = M_{ijk} + R_{ij} (\phi_{ijk}/\phi_{ij+}) + S_k (\phi_{ijk}/\phi_{++k}) \qquad 3.23$$

and so the E step is:

$$T^{(p)}_{ijk} = M_{ijk} + R_{ij} (\phi^{(p)}_{ijk}/\phi^{(p)}_{ij+}) + S_k (\phi^{(p)}_{ijk}/(\phi^{(p)}_{++k})$$

The M step calls for finding the maximum likelihood estimate of ϕ_{ijk} given $T_{ijk} = T^{(p)}_{ijk}$. This estimate will then be used for $\phi^{(p+1)}_{ijk}$.

If a saturated model is being considered then the maximum likelihood estimate of ϕ_{ijk} given T_{ijk} is just T_{ijk} and so for this case

$$\phi^{(p+1)}_{ijk} = T^{(p)}_{ijk}$$

If an unsaturated hierarchical log linear model is being considered then the log linear modelling process is applied to $T^{(p)}_{ijk}$ and the resulting cell estimates are used for $\phi^{(p+1)}_{ijk}$. Note that the expected values of cell counts

*This follows from the properties of the Poisson or the multinomial distribution

(i.e. ϕ_{ijk}'s) are not the usual way of expressing the parameters of a log linear model but they are in one-to-one correspondence with the usual parameters (such as $U_{12(ij)}$ etc.). Consequently it is proper to consider the parameters of the log linear model to be the expected values of the cell counts. According to the process of log linear modelling $\phi_{ijk}^{(p+1)}$ is determined by forming the appropriate margins of $T_{ijk}^{(p)}$ (minimal sufficient statistics for ϕ_{ijk}) and performing IPF with these margins and a core of all ones, (the procedure is described in some detail in Section 3.2.2).
This completes the outline of how the EM algorithm works in a typical case of interest.

As with IPF there are computational questions of interest regarding EM:

1. Under what circumstances does the process converge and under what circumstances does it not converge.

2. Is there a unique result independent of the initialization of $\phi_{ijk}^{(o)}$?

3. When does the process converge to a maximum likelihood solution?

These questions appear to be hard to answer in general. Dempster et al stated, "we demonstrate.....the key results which assert that successive iterations always increase the likelihood, and that convergence implies a stationary point of the likelihood." This statement falls far short of guaranteeing that the process always converges to a unique point of maximum likelihood. Such a guarantee is hardly to be expected. It is easy to devise examples which do not converge to a unique solution and in which there is no unique maximum likelihood solution.

The next section considers questions of convergence and uniqueness in typical examples such as would be encountered in dealing with tables of incompletely classified count data which are to be described by log linear models using the EM method.

3.4.3 Convergence and Uniqueness in Typical Examples

Given two models, one of which is a direct extension in complexity of the other, i.e. such that the simple model is equivalent to the complex one with particular values (e.g. zero) for some of its parameters, then it is clear that the simpler model will lead to fewer local maxima of the likelihood than the more complex model (or at most the same number). It seems that similar statements should hold in regard to the EM procedure: it might be expected that the EM procedure would be more likely to yield a unique result when applied to a simpler model than when applied to a more complex model. In particular if the EM algorithm is applied to a particular situation assuming on the one hand a saturated or high order log linear model, and on the other assuming a lower order log linear model, then uniqueness of the result would be more likely in the latter case. This conjecture is based on intuition and a very limited amount of experience with the procedure.

Initial experience with the EM procedure applied to saturated models suggests that it usually converges but sometimes very slowly. EM appears to converge much more slowly than IPF. Non-uniqueness of the EM method may be confined to the case where the completely classified data matrix has zero cells.

Although it is not possible to answer questions of convergence and uniqueness in general, the conjectures stated above may be supplemented by a more detailed analysis of a special case.

Consider a case where a completely classified data matrix M_{ijk} is available as well as one dimensional data matrices A_i, B_j, C_k. (The dimensionality of M_{ijk} is arbitrary but the imcompletely classified matrices must be one dimensional for this analysis.) Suppose that it is desired to fit a saturated log linear model to this data using the EM method. The EM calculation proceeds as follows:

$$T_{ijk}^{n+1} = M_{ijk} + (A_i/T_{i++}^{(n)})\, T_{ijk}^{(n)} + (B_j/T_{+j+}^{(n)})\, T_{ijk}^{(n)} + (C_k/T_{++k}^{(n)})\, T_{ijk}^{(n)}$$

If this process converges, the result $T_{ijk} = \lim_{n\to\infty} T_{ijk}^{(n)}$ satisfies the following equations:

$$T_{ijk} = M_{ijk} + (A_i/T_{i++})\, T_{ijk} + (B_j/T_{+j+})\, T_{ijk} + (C_k/T_{++k})\, T_{ijk}$$

or

$$T_{ijk}\,(1 - A_i/T_{i++} - B_j/T_{+j+} - C_k/T_{++k}) = M_{ijk} \qquad 3.24$$

Now suppose that $A_+, B_+, C_+ >> M_{+++}$ i.e. the core, M_{ijk}, has small cell counts compared to the marginal data A_i, B_j, C_k. This represents a situation which is expected to lead to a greater lack of determinism in the result than any other case (if enough completely classified data is available then there should be no indeterminism). Assuming further that M_{ijk} is small compared to T_{ijk} it follows that

$$(1 - A_i/T_{i++} - B_j/T_{+j+} - C_k/T_{++k}) \approx 0.$$

This implies that*

$$T_{i++} \approx \alpha\, A_i$$

$$T_{+j+} \approx \beta\, B_j$$

$$T_{++k} \approx \gamma\, C_k$$

with

$$1/\alpha + 1/\beta + 1/\gamma = 1$$

and

$$\alpha\, A_+ = \beta\, B_+ = \gamma\, C_+$$

*For example, A_i/T_{i++} approximately depends only on j and k and not on i and so is approximately a constant, α.

At the same time from equation 3.24 T_{ijk} has the form**

$$T_{ijk} = M_{ijk}/(U_i + V_j + W_k)$$

Here U_i, V_j, W_k are determined by the marginal conditions.

A resemblance to formulation 1 of IPF may be observed: (replacing $\approx$ by =).

Find T_{ijk} of the form:

$$T_{ijk} = M_{ijk}/(U_i + V_j + W_k)$$

(M_{ijk} given, U_i, V_j, W_k to be found)

such that

$$T_{i++} = a_i$$

$$T_{+j+} = b_j$$

$$T_{++k} = c_k$$

here

$$a_i = \alpha A_i = ((A_+ + B_+ + C_+)/A_+)\ A_i = ((G/A_+)A_i), b_j = (G/B_+)\ B_j,\ c_j = (G/C_+)\ C_j$$

This can even be given an equivalent math programming formulation (if $M_{ijk} > 0$):

$$\underset{T_{ijk}}{\text{Max}} \sum_{rst} (M_{rst} \log_e (T_{rst}/M_{rst}) - T_{rst}) \qquad 3.25$$

**$U_i = A_i/T_{i++} + d_1$, $V_j = B_j/T_{+j+} + d_2$, $W_k = C_k/T_{++k} + d_3$, where d_1, d_2, d_3 are arbitrary except $d_1 + d_2 + d_3 = 1$.

subject to

$$T_{i++} = a_i$$
$$T_{+j+} = b_j$$
$$T_{++k} = c_k \qquad 3.26$$

and

$$T_{ijk} \geq 0$$

As in the case of IPF, existence and uniqueness of a solution T_{ijk} can be guaranteed if and only if the feasibility conditions 3.26 are satisfied. It is easy to show that the feasibility conditions always have a solution. (This comment is restricted to the case of one dimensional margins.)

If $M_{ijk} = 0$ for any ijk then, from the form $T_{ijk} = M_{ijk}/(U_i+V_j+W_k)$ there is a threat of indeterminateness of the solution. If too many of the $M_{ijk}=0$ it seems likely that no unique solution for T_{ijk} will exist. (The math programming formulation ensures uniqueness but does not apply if $M_{ijk} = 0$).

However, if $M_{ijk} > 0$ for all ijk then the EM process can have only one solution and there is no possibility of the solution depending on the initialization of $T_{ijk}^{(0)}$. The existence of the solution to 3.25, 3.26 i.e., to the alternative formulation is guaranteed, but this does not strictly prove that the EM process itself converges. However, it seems likely that it will.

In this simple case with non-zero values in all cells of M_{ijk}, using EM with a saturated model cannot lead to more than one solution (i.e. depending on the initialization). Consequently the previous conjecture that simpler models (i.e., models with fewer parameters) are more likely to lead to unique solutions would suggest that all log linear models would lead to a unique solution if the process converges. It seems likely that the process converges and that the unique solution is the maximum likelihood solution. These arguments strictly speaking apply only to the case where the margins dominate the core and where the margins are all one dimensional.

With a large sample determining M_{ijk} and with higher dimensional margins given for the incompletely classified data, it might be conjectured that non uniqueness would be less likely (than with small M_{ijk} and one dimensional margins). As a consequence it is conjectured that the EM method applied as outlined in Section 3.4.2 will always lead to a unique maximum likelihood estimate of a log linear model if it converges unless some cells of the completely classified data are empty. In the case of margins of higher dimensions it is possible that convergence depends on the feasibility of the margins (no conjecture is offered in the case of infeasible margins but feasibility of the margins may not matter in the case of EM).

It is worth noting again that the conjectures concerning convergence in this section are based on limited experience and on the above analysis of a special case.

3.5 PROPERTIES OF TECHNIQUES IN APPLICATIONS

This section provides some brief observations on the implications of the properties of the techniques discussed in this report in regard to the application of these techniques.

3.5.1 IPF

The IPF technique is for the most part motivated by considerations not usually considered to be a part of classical statistics. Since the data marginal matrices completely determine the margins of the result, IPF should be used only if the marginal matrices to be used are considered reliable representatives of the population to be represented by the result. This means that they (each) should have a sample size large enough to be considered statistically stable and that they should be based on samples representative of the population of interest.

On the other hand, the data core need not be representative of the population of interest in its lower order interactions (those characteristics of the variable combinations represented in the marginal data) but it completely determines the higher order interactions. Since it is usually a straightforward matter to decide if marginal data is representative of the population of interest, the most problematic decision in connection with IPF is whether the core is appropriate.

If more than one core is available each with different margins but (possibly) with the same higher order interactions (e.g. multivariate data from sources with different time and space frames) then the higher order interactions of each can be extracted (using log linear modelling techniques) and compared to validate the assumption of equality of higher order interactions. Also the sensitivity of the result with respect to variations in the core can be determined.

The statistical stability of the core is as important as its representativeness. Formulation 1 shows that zeroes in the core are preserved

in the result. The same is true of thin spots. It also is clear that the result of IPF is sensitive on a cell by cell basis to fluctuations in cell counts (a p% change in a low population cell dominated by the rows and columns it is in will result in a p% change in the result). This shows that statistical stability of the core is important. A thinly populated core should never be subjected to IPF with high population margins with the hopes that the proces will "smooth out" the thin spots in the core. Instead, the core can in some cases be smoothed out prior to applying IPF by first using it to construct a log linear model with some of the highest order interactions left out. However, some interactions not represented in the marginal data matrices must be retained in the core or else the IPF process will yield the same result as would be obtained using an independent core (i.e. a core of all ones; See Section 3.3.3).

3.5.2 Log Linear Modelling

Log linear modelling is used to provide smoothed cell estimates of multivariate data and to separate interactions or multivariable effects so that they may be evaluated and compared. Log linear models help determine which interactions are significant in the data and must be considered in further analysis or in using the data in further constructions.

It also allows smoother and therefore lower variance cell estimates to be produced. This can be of use in connection with IPF.

3.5.3 EM

The EM method is primarily of use for combining data from several samples each classified by a different set of variables but each from the same population. In the case where no external "core" or measure of higher order interactions is available, the EM method provides a smooth and even "optimal"* method for combining the available marginal (and completely classified if available) data.

*If it is assumed that the various data sets represent the same population and that the data are multinomially distributed according to a log linear model, then the EM method leads to maximum likelihood estimates. In this sense it is "optimal."

The key assumption is that all data is from the same population. It is advisable to statistically test this assumption (chi square tests comparing the individual data sets to estimates of them derived from the complete data estimated by EM can be used).

In cases where several good, incompletely classified data sets each pertinent to (and representative of) the population of interest are available, the choice of using EM or IPF to combine them depends on the following considerations:

1. If a core from a separate population is to be used to determine the higher order interactions, then IPF must be used.

2. If no external core is to be used and especially if an internal completely classified data set is available, then EM should be used.

3 In spite of its theoretically optimal properties, the EM method is rather sharply limited by the requirement that all data be assumed to be from the same population.

In summary IPF is capable of producing rather dramatic extensions of input data while the EM method bolsters the effective sample size of a data set by combining with an incompletely classified data set.

3.6 SUMMARY

This chapter has considered three analytical techniques:

1. Log linear modelling,
2. Iterative proportional fitting,
3. The EM Expectation-Maximization) algorithm,

focussing on aspects likely to be of interest in producing useful estimates of multivariate exposure to driving accidents.

The section on log linear modelling gave a brief overview of that process. The concepts and procedures are also of use in regard to the other two methods. The last subsection of the section on log linear modelling considered the problems involved when log linear modelling is to be applied to continuous data (such as VMT). It is concluded that log linear models are of great interest for modelling continuous data such as VMT. The main problems are with the means of estimating the model parameters and with assessing the goodness of fit.

It is shown that the standard method for estimating descrete log linear models (based on IPF) is still satisfactory for the continuous case (although it does not necessarily yield maximum likelihood estimates) but the chi square statistics (calculated as if the cell VMT totals were cell counts) are not in general valid (i.e., not chi square distributed with the appropriate degrees of freedom). However, if certain conditions on the data are valid, the formal chi square statistic is easily scaled to a valid chi square (distributed) statistic (with the usual degrees of freedom). The conditions except for one are mild and probably satisfied in very many instances. The condition which will often not hold is that the ratio of VMT mean to VMT variance is constant across cells. This condition would usually hold only if the variance in cell VMT was mostly caused by fluctuations in the number of records in each cell (i.e., if each record made about the same contribution to VMT).

Even though the condition is not reasonable in many cases, the scale factor for the chi square statistic can be calculated. A goodness of fit could also be based on some robust procedure such as jackknife, infinitesimal jackknife or sample splitting unless and until the adjusted chi square statistic proposed here is found to work satisfactorily.

The section on IPF describes the procedure for calculating IPF and then goes on to show two other equivalent formulations leading to the same solution. The alternative formulations are not alternative methods of solution but formulations of problems which have the same solution as the usual formulation of IPF. From the alternative formulations various properties of the IPF procedure are developed:

1. The solution, if it exists, is unique.

2. The resulting matrix is as close to the core matrix as possible (according to an information type metric which is concerned with the ratio of the result matrix to the core matrix) subject to the marginal and positivity constraints.

3. The result of IPF has the same margin-absent effects as the core. The margin-absent effects are defined in Section 3.2.2. They are roughly speaking the higher order interactions.

4. The result of IPF is unaffected by the margin-present effects in the core.

5. IPF has a solution only if the margins are feasible.

The problem of the feasibility of the margins is discussed in Section 3.3 and also in Appendix B. It is concluded that the best practical method for determining margin feasbility is to apply IPF and observe the assymptotic properties (convergence or cycling, development of zeroes or not). The application of IPF in reasonable situations should not often lead to infeasible margins but the discussion given here should assist in handling any such cases that might arise.

The section on the EM method gives a brief introduction to the method and then presents the method as it would apply to developing log linear models for incompletely classified data. A discussion on convergence and uniqueness is given based on an explicit analysis of a case which might be expected to have a greater propensity for indeterminacy than most other cases: saturated modelling in the case of one dimensional marginal data (incompletely classified data sets) and a thinly populated completely classified data set. In this special case it is concluded that if the process converges the answer will be unique so long as the completely classified data set does not have too many zero cells. Extrapolating to other cases it is conjectured that convergence almost always occurs (if the margins are fairly compatible) and that uniqueness is guaranteed if there is a completely classified data set with no zero cells. If there is no completely classified data set then the process is likely to converge to a unique maximum likelihood estimate only if the order of the log linear model is not too high.

The EM method appears to be very slowly convergent in some cases.

Since the EM method is derived for the case where all the data sets come from populations having the same distribution its application may be rather restricted.

4. COMPUTATIONAL EXAMPLES

In this section, several of the techniques described in Section 3 are applied to exposure data sets of modest complexity. One of the major benefits of these exercises was to test out software, develped for the TSC computer (a PDP10/KL), to compute margins of general arrays, to perform Iterative Proportional Fitting (IPF) on a broad class of cores and margins, and to compute log linear models for a large class of arrays. These computer programs are listed in Appendix C. The set of problems that can be addressed is limited only by computer storage requirements. Although not represented by an example in this section, the software has been further extended to compute maximum likelihood cell-count estimates for a variety of cases involving incomplete data, via the EM algorithm.

In Section 4.1, a variety of log linear models are fit to a 4-dimensional Vehicle Miles Travelled (VMT) data set derived from the 1977 Nationwide Personal Transportation Survey (NPTS). Goodness of fit measures are given for first, second and third order models. (See Section 3 for a description of the special problems involved in applying log linear theory to VMT (continuous) data.)

In Section 4.2, an example of IPF with a dummy core of 1's is given. This is an elementary example of incomplete data classification discussed in Section 2.2. Ten bivariate margins are given representing pairwise distribution of VMT among five variables, taken from a North Carolina driver survey from which an estimate of the fully classified 5-dimensional array is made by fitting the margins to a 5-dimensional core of ones. Three-dimensional margins from the estimated array are compared with actual three-way distributions taken from the driver survey, as an indication of effectiveness of the estimation.

In Section 4.3, an example of IPF with a real core is given. This is an elementary example of the situation described in Section 2.8 - the updating of out-of-date data. Matrices of registered drivers by age and sex categories for 1975 and for 1979 are each fit to the corresponding 1980 margins by IPF.

The results are compared with the actual 1980 matrix. Since, in this case the data are counted data, a statistical goodness of fit test is applied. Saturated log linear model coefficients are calculated for the data sets, to trace the effect of IPF on the interactions. A major theoretical result of Section 3.3.3 is demonstrated computationally.

4.1 LOG LINEAR MODELLING - NPTS DATA SET APPLICATIONS

The log linear modelling method was applied to a multivariate data set consisting of 1977 annual VMT, classified by four variables: driver age, driver sex, vehicle weight and model year. The definition of variable levels (together with a key to their use with data tables) is shown in Table 2. The actual data set is displayed in Table 3. The data set was aggregated from a larger multivariate set (more variables and more levels), constructed by TSC from the 1977 NPTS data tapes. The particular 4-variable set used here was chosen for several reasons, the principal one being to have a relatively smooth data set of modest size, for an initial software testing. (The more disaggregate data are quite noisy and have many zero cells.)

Three hierarchical log linear models were fit to the data, using the standard technique described in Section 3.2.

The first model fit was a homogeneous hierarchical model of order 3 (meaning all 3-variable effects and lower order effects were present, all effects for more than 3 variables were not present). The four 3-dimensional margins used as sufficient statistics for the estimation are shown in Tables 4, 5, 6, and 7. The outcome cell estimates for this model are shown in Table 8.

The second model fit was a homogeneous hierarchical model of order 2. The six 2-dimensional margins (sufficient statistics) are shown in Tables 9 and 10. The cell estimates under this model are shown in Table 11.

The final log linear model fitted was a homogenous hierarchical model of order 1. The four 1-dimensional margins (sufficient statistics) are shown in Table 12. The cell estimates under this model are shown in Table 13.

TABLE 2. DEFINITION OF VARIABLE LEVELS

VARIABLE/LEVELS	KEY
SEX OF DRIVER	
Male	M
Female	F
AGE OF DRIVER	
0-24	0+
25-34	25+
35-44	35+
45-54	45+
55 and older	55+
WEIGHT OF VEHICLE	
0-2500 lbs.	0.0+
2501-3500 lbs.	2.5+
3501-4500 lbs.	3.5+
over 4500 lbs.	4.5+
YEAR OF MODEL	
1976-1978	76+
1973-1975	73+
1969-1972	69+
1965-1968	65+
1964 or older	0+

TABLE 3. 4-WAY TABLE OF VMT (BILLIONS)

DRIVER AGE	DRIVER SEX	VEHICLE WGT.	VEHICLE YR.	VMT BILLION
0+	M	0.0+	76+	2.339
0+	F	0.0+	76+	2.338
0+	M	2.5+	76+	7.024
0+	F	2.5+	76+	5.251
0+	M	3.5+	76+	11.344
0+	F	3.5+	76+	4.634
0+	M	4.5+	76+	2.097
0+	F	4.5+	76+	0.601
0+	M	0.0+	73+	4.376
0+	F	0.0+	73+	3.748
0+	M	2.5+	73+	11.202
0+	F	2.5+	73+	6.889
0+	M	3.5+	73+	15.734
0+	F	3.5+	73+	8.332
0+	M	4.5+	73+	1.546
0+	F	4.5+	73+	1.057
0+	M	0.0+	69+	3.454
0+	F	0.0+	69+	1.908
0+	M	2.5+	69+	11.758
0+	F	2.5+	69+	8.229
0+	M	3.5+	69+	12.561
0+	F	3.5+	69+	6.447
0+	M	4.5+	69+	1.441
0+	F	4.5+	69+	1.199
0+	M	0.0+	65+	0.832
0+	F	0.0+	65+	0.118
0+	M	2.5+	65+	7.668
0+	F	2.5+	65+	3.135
0+	M	3.5+	65+	6.383
0+	F	3.5+	65+	2.960
0+	M	4.5+	65+	0.879
0+	F	4.5+	65+	0.490
0+	M	0.0+	0+	0.566
0+	F	0.0+	0+	0.465
0+	M	2.5+	0+	2.313
0+	F	2.5+	0+	0.869
0+	M	3.5+	0+	3.082
0+	F	3.5+	0+	0.916
0+	M	4.5+	0+	0.199
0+	F	4.5+	0+	0.048
25+	M	0.0+	76+	6.112
25+	F	0.0+	76+	3.021
25+	M	2.5+	76+	12.814
25+	F	2.5+	76+	5.930
25+	M	3.5+	76+	20.959
25+	F	3.5+	76+	7.633
25+	M	4.5+	76+	1.996
25+	F	4.5+	76+	0.605
25+	M	0.0+	73+	8.145
25+	F	0.0+	73+	2.874
25+	M	2.5+	73+	17.715
25+	F	2.5+	73+	8.501
25+	M	3.5+	73+	22.044
25+	F	3.5+	73+	11.162
25+	M	4.5+	73+	3.507
25+	F	4.5+	73+	2.138
25+	M	0.0+	69+	2.945
25+	F	0.0+	69+	1.659
25+	M	2.5+	69+	18.521
25+	F	2.5+	69+	9.782
25+	M	3.5+	69+	16.542
25+	F	3.5+	69+	8.378
25+	M	4.5+	69+	3.552
25+	F	4.5+	69+	0.927
25+	M	0.0+	65+	1.625
25+	F	0.0+	65+	0.247
25+	M	2.5+	65+	7.441
25+	F	2.5+	65+	2.154
25+	M	3.5+	65+	7.200
25+	F	3.5+	65+	3.158
25+	M	4.5+	65+	1.542
25+	F	4.5+	65+	0.678
25+	M	0.0+	0+	0.753
25+	F	0.0+	0+	0.125
25+	M	2.5+	0+	2.885
25+	F	2.5+	0+	0.383
25+	M	3.5+	0+	3.296
25+	F	3.5+	0+	0.589
25+	M	4.5+	0+	0.447
25+	F	4.5+	0+	0.093
35+	M	0.0+	76+	5.854
35+	F	0.0+	76+	1.388
35+	M	2.5+	76+	9.553
35+	F	2.5+	76+	3.436
35+	M	3.5+	76+	13.758
35+	F	3.5+	76+	6.610
35+	M	4.5+	76+	2.655
35+	F	4.5+	76+	1.218
35+	M	0.0+	73+	2.342
35+	F	0.0+	73+	1.832
35+	M	2.5+	73+	11.094
35+	F	2.5+	73+	5.855
35+	M	3.5+	73+	15.139
35+	F	3.5+	73+	8.546
35+	M	4.5+	73+	4.328
35+	F	4.5+	73+	1.559
35+	M	0.0+	69+	3.644
35+	F	0.0+	69+	1.539
35+	M	2.5+	69+	9.864
35+	F	2.5+	69+	4.194
35+	M	3.5+	69+	16.734
35+	F	3.5+	69+	6.819
35+	M	4.5+	69+	1.993
35+	F	4.5+	69+	1.993
35+	M	0.0+	65+	0.449
35+	F	0.0+	65+	0.645
35+	M	2.5+	65+	4.904
35+	F	2.5+	65+	2.376
35+	M	3.5+	65+	6.127
35+	F	3.5+	65+	1.926
35+	M	4.5+	65+	1.334
35+	F	4.5+	65+	0.515
35+	M	0.0+	0+	0.857
35+	F	0.0+	0+	0.040
35+	M	2.5+	0+	1.946
35+	F	2.5+	0+	0.452
35+	M	3.5+	0+	2.181
35+	F	3.5+	0+	0.239
35+	M	4.5+	0+	0.752
35+	F	4.5+	0+	0.188
45+	M	0.0+	76+	2.927
45+	F	0.0+	76+	0.715
45+	M	2.5+	76+	9.850
45+	F	2.5+	76+	4.454
45+	M	3.5+	76+	15.372
45+	F	3.5+	76+	4.903
45+	M	4.5+	76+	2.725
45+	F	4.5+	76+	1.363
45+	M	0.0+	73+	2.833
45+	F	0.0+	73+	0.671
45+	M	2.5+	73+	10.379
45+	F	2.5+	73+	5.925
45+	M	3.5+	73+	16.636
45+	F	3.5+	73+	6.715
45+	M	4.5+	73+	4.188
45+	F	4.5+	73+	2.014
45+	M	0.0+	69+	1.427
45+	F	0.0+	69+	1.643
45+	M	2.5+	69+	9.919
45+	F	2.5+	69+	4.071
45+	M	3.5+	69+	12.061
45+	F	3.5+	69+	3.886
45+	M	4.5+	69+	2.811
45+	F	4.5+	69+	1.120
45+	M	0.0+	65+	0.357
45+	F	0.0+	65+	0.421
45+	M	2.5+	65+	4.123
45+	F	2.5+	65+	2.171
45+	M	3.5+	65+	5.392
45+	F	3.5+	65+	2.504
45+	M	4.5+	65+	0.557
45+	F	4.5+	65+	0.536
45+	M	0.0+	0+	0.152
45+	F	0.0+	0+	0.074
45+	M	2.5+	0+	1.528
45+	F	2.5+	0+	0.325
45+	M	3.5+	0+	1.878
45+	F	3.5+	0+	0.503
45+	M	4.5+	0+	0.120
45+	F	4.5+	0+	0.058
55+	M	0.0+	76+	1.843
55+	F	0.0+	76+	0.588
55+	M	2.5+	76+	7.416
55+	F	2.5+	76+	2.420
55+	M	3.5+	76+	15.239
55+	F	3.5+	76+	4.801
55+	M	4.5+	76+	2.142
55+	F	4.5+	76+	0.609
55+	M	0.0+	73+	2.044
55+	F	0.0+	73+	0.684
55+	M	2.5+	73+	9.324
55+	F	2.5+	73+	4.158
55+	M	3.5+	73+	14.225
55+	F	3.5+	73+	4.940
55+	M	4.5+	73+	5.041
55+	F	4.5+	73+	1.154
55+	M	0.0+	69+	2.231
55+	F	0.0+	69+	0.798
55+	M	2.5+	69+	8.232
55+	F	2.5+	69+	3.207
55+	M	3.5+	69+	15.011
55+	F	3.5+	69+	3.815
55+	M	4.5+	69+	4.020
55+	F	4.5+	69+	0.718
55+	M	0.0+	65+	1.531
55+	F	0.0+	65+	0.259
55+	M	2.5+	65+	4.379
55+	F	2.5+	65+	1.371
55+	M	3.5+	65+	5.036
55+	F	3.5+	65+	2.213
55+	M	4.5+	65+	1.125
55+	F	4.5+	65+	0.342
55+	M	0.0+	0+	0.526
55+	F	0.0+	0+	0.126
55+	M	2.5+	0+	0.913
55+	F	2.5+	0+	0.361
55+	M	3.5+	0+	1.160
55+	F	3.5+	0+	0.281
55+	M	4.5+	0+	0.177
55+	F	4.5+	0+	0.190

SOURCE: 1977 NATIONWIDE PERSONAL TRANSPORTATION STUDY

TABLE 4. THREE-DIMENSIONAL MARGIN: AGE, SEX, WEIGHT OF VEHICLE
(Annual Billions of VMT, 1977 NPTS)

AGE	SEX	
	MALE	FEMALE
0-24	11.606	8.577
25-34	19.575	7.926
35-44	13.146	5.436
45-54	7.696	3.525
55 & over	8.235	2.455

Weight of Vehicle <2500 lbs.

AGE	SEX	
	MALE	FEMALE
0-24	39.954	24.373
25-34	59.175	26.746
35-44	37.362	16.314
45-54	35.799	16.946
55 & over	30.264	11.517

Weight of Vehicle 2501-3500 lbs.

AGE	SEX	
	MALE	FEMALE
0-24	49.104	23.258
25-34	70.041	31.520
35-44	53.936	24.131
45-54	51.335	18.512
55 & over	50.670	16.050

Weight of Vehicle 3501-4500 lbs.

AGE	SEX	
	MALE	FEMALE
0-24	6.161	4.195
25-34	10.044	4.350
35-44	10.353	5.473
45-54	10.393	5.091
55 & over	12.506	3.012

Weight of Vehicle over 4500 lbs.

TABLE 5. THREE-DIMENSIONAL MARGIN: AGE, SEX, YEAR OF MODEL
(Annual Billions of VMT, 1977 NPTS)

AGE	SEX	
	MALE	FEMALE
0-24	22.800	12.824
25-34	41.681	17.190
35-44	31.816	12.652
45-54	30.875	11.434
55 & over	26.640	8.418

Year of Model 1976-78

AGE	SEX	
	MALE	FEMALE
0-24	32.857	20.797
25-34	51.411	24.673
35-44	32.895	17.792
45-54	34.028	15.326
55 & over	30.634	10.936

Year of Model 1973-75

AGE	SEX	
	MALE	FEMALE
0-24	29.255	17.782
25-34	40.559	21.346
35-44	32.235	14.528
45-54	26.218	10.721
55 & over	29.553	8.538

Year of Model 1969-72

AGE	SEX	
	MALE	FEMALE
0-24	15.753	6.703
25-34	17.803	6.237
35-44	12.815	5.462
45-54	10.430	5.632
55 & over	12.071	4.185

Year of Model 1965-68

AGE	SEX	
	MALE	FEMALE
0-24	6.160	2.297
25-34	7.381	1.097
35-44	5.037	0.920
45-54	3.673	0.960
55 & over	2.775	0.958

Year of Model 1964 or older

TABLE 6. THREE-DIMENSIONAL MARGIN: AGE, WEIGHT OF VEHICLE AND YEAR OF MODEL
(Annual Billions of VMT, 1977 NPTS)

Age of Driver 0-24

YEAR OF MODEL	Weight of Vehicle (1000 lbs.) 0-2.5	2.5-3.5	3.5-4.5	over 4.5
76-78	4.672	12.271	15.978	2.697
73-75	8.124	18.091	24.036	3.403
69-72	5.401	19.988	19.008	2.640
65-68	0.950	10.795	9.342	1.369
64 or older	1.031	3.182	3.998	0.247

Age of Driver 25-34

YEAR OF MODEL	Weight of Vehicle (1000 lbs.) 0-2.5	2.5-3.5	3.5-4.5	over 4.5
76-78	9.134	18.543	28.592	2.601
73-75	11.019	26.215	33.206	5.645
69-72	4.604	28.303	25.520	3.478
65-68	1.866	9.595	10.358	2.220
64 or older	0.878	3.265	3.883	0.450

Age of Driver 35-44

YEAR OF MODEL	Weight of Vehicle (1000 lbs.) 0-2.5	2.5-3.5	3.5-4.5	over 4.5
76-78	7.242	12.989	20.364	3.873
73-75	4.174	16.949	23.685	5.879
69-72	5.174	14.058	23.545	3.986
65-68	1.094	7.281	8.053	1.848
64 or older	0.897	2.399	2.421	0.240

Age of Driver 45-54

YEAR OF MODEL	Weight of Vehicle (1000 lbs.) 0-2.5	2.5-3.5	3.5-4.5	over 4.5
76-78	3.642	14.304	20.275	4.088
73-75	3.504	16.305	23.351	6.194
69-72	3.070	13.990	15.947	3.931
65-68	0.778	6.295	7.896	1.093
64 or older	0.226	1.852	2.377	0.178

Age of Driver over 55

YEAR OF MODEL	Weight of Vehicle (1000 lbs.) 0-2.5	2.5-3.5	3.5-4.5	over 4.5
76-78	2.431	9.836	20.040	2.751
73-75	2.728	13.483	19.164	6.196
69-72	3.089	11.439	18.826	4.738
65-68	1.790	5.750	7.249	1.467
64 or older	0.652	1.274	1.441	0.366

TABLE 7. THREE-DIMENSIONAL MARGIN: SEX OF DRIVER, WEIGHT OF VEHICLE, AND YEAR OF MODEL (Annual Billions of VMT, 1977 NPTS)

Male Drivers				
	Weight of Vehicle (1000 lbs.)			
YEAR OF MODEL	0-2.5	2.5-3.5	3.5-4.5	over 4.5
76-78	19.075	46.453	76.669	11.615
73-75	19.740	59.715	83.778	18.594
69-72	13.800	58.294	72.909	12.817
65-68	4.789	28.507	30.138	5.437
64 or older	2.853	9.585	11.593	0.994

Female Drivers				
	Weight of Vehicle (1000 lbs.)			
YEAR OF MODEL	0-2.5	2.5-3.5	3.5-4.5	over 4.5
76-78	8.052	21.490	28.581	4.395
73-75	9.808	31.327	39.665	8.723
69-72	7.538	24.483	29.937	5.957
65-68	1.691	11.208	12.760	2.559
64 or older	0.830	2.387	2.529	0.487

TABLE 8. CELL ESTIMATES FROM LONG LINEAR MODEL BASED ON 3-DIMENSIONAL MARGINS

DRIVER AGE	DRIVER SEX	VEHICLE WGT.	VEHICLE YR.	VMT BILLION
0+	M	0.0+	76+	2.736
0+	F	0.0+	76+	1.941
0+	M	2.5+	76+	7.422
0+	F	2.5+	76+	4.849
0+	M	3.5+	76+	10.944
0+	F	3.5+	76+	5.034
0+	M	4.5+	76+	1.638
0+	F	4.5+	76+	0.999
0+	M	0.0+	73+	4.647
0+	F	0.0+	73+	3.476
0+	M	2.5+	73+	10.663
0+	F	2.5+	73+	7.429
0+	M	3.5+	73+	15.577
0+	F	3.5+	73+	8.459
0+	M	4.5+	73+	1.974
0+	F	4.5+	73+	1.433
0+	M	0.0+	69+	2.911
0+	F	0.0+	69+	2.499
0+	M	2.5+	69+	12.007
0+	F	2.5+	69+	7.981
0+	M	3.5+	69+	12.035
0+	F	3.5+	69+	6.174
0+	M	4.5+	69+	1.502
0+	F	4.5+	69+	1.138
0+	M	0.0+	65+	0.836
0+	F	0.0+	65+	0.314
0+	M	2.5+	65+	7.577
0+	F	2.5+	65+	3.218
0+	M	3.5+	65+	6.677
0+	F	3.5+	65+	2.665
0+	M	4.5+	65+	0.863
0+	F	4.5+	65+	0.505
0+	M	0.0+	0+	0.675
0+	F	0.0+	0+	0.356
0+	M	2.5+	0+	2.286
0+	F	2.5+	0+	0.896
0+	M	3.5+	0+	3.072
0+	F	3.5+	0+	0.926
0+	M	4.5+	0+	0.127
0+	F	4.5+	0+	0.120
25+	M	0.0+	76+	6.813
25+	F	0.0+	76+	2.521
25+	M	2.5+	76+	12.886
25+	F	2.5+	76+	5.657
25+	M	3.5+	76+	20.254
25+	F	3.5+	76+	8.338
25+	M	4.5+	76+	1.928
25+	F	4.5+	76+	0.674
25+	M	0.0+	73+	7.779
25+	F	0.0+	73+	3.240
25+	M	2.5+	73+	17.651
25+	F	2.5+	73+	8.564
25+	M	3.5+	73+	22.083
25+	F	3.5+	73+	11.123
25+	M	4.5+	73+	1.898
25+	F	4.5+	73+	1.746
25+	M	0.0+	69+	2.973
25+	F	0.0+	69+	1.630
25+	M	2.5+	69+	18.483
25+	F	2.5+	69+	9.840
25+	M	3.5+	69+	16.859
25+	F	3.5+	69+	8.660
25+	M	4.5+	69+	2.263
25+	F	4.5+	69+	1.215
25+	M	0.0+	65+	1.446
25+	F	0.0+	65+	0.421
25+	M	2.5+	65+	7.308
25+	F	2.5+	65+	2.288
25+	M	3.5+	65+	7.442
25+	F	3.5+	65+	2.916
25+	M	4.5+	65+	1.607
25+	F	4.5+	65+	0.613
25+	M	0.0+	0+	0.764
25+	F	0.0+	0+	0.114
25+	M	2.5+	0+	2.867
25+	F	2.5+	0+	0.397
25+	M	3.5+	0+	3.102
25+	F	3.5+	0+	0.403
25+	M	4.5+	0+	0.348
25+	F	4.5+	0+	0.102
35+	M	0.0+	76+	5.257
35+	F	0.0+	76+	1.985
35+	M	2.5+	76+	9.173
35+	F	2.5+	76+	3.916
35+	M	3.5+	76+	14.634
35+	F	3.5+	76+	5.729
35+	M	4.5+	76+	2.752
35+	F	4.5+	76+	1.121
35+	M	0.0+	73+	2.833
35+	F	0.0+	73+	1.341
35+	M	2.5+	73+	11.097
35+	F	2.5+	73+	5.852
35+	M	3.5+	73+	15.288
35+	F	3.5+	73+	8.398
35+	M	4.5+	73+	3.677
35+	F	4.5+	73+	2.202
35+	M	0.0+	69+	3.497
35+	F	0.0+	69+	1.678
35+	M	2.5+	69+	9.722
35+	F	2.5+	69+	4.336
35+	M	3.5+	69+	16.449
35+	F	3.5+	69+	7.995
35+	M	4.5+	69+	2.568
35+	F	4.5+	69+	1.419
35+	M	0.0+	65+	0.835
35+	F	0.0+	65+	0.289
35+	M	2.5+	65+	5.317
35+	F	2.5+	65+	1.963
35+	M	3.5+	65+	5.502
35+	F	3.5+	65+	2.551
35+	M	4.5+	65+	1.190
35+	F	4.5+	65+	0.658
35+	M	0.0+	0+	0.754
35+	F	0.0+	0+	0.143
35+	M	2.5+	0+	2.052
35+	F	2.5+	0+	0.346
35+	M	3.5+	0+	2.861
35+	F	3.5+	0+	0.358
35+	M	4.5+	0+	0.167
35+	F	4.5+	0+	0.073
45+	M	0.0+	76+	2.582
45+	F	0.0+	76+	1.060
45+	M	2.5+	76+	9.897
45+	F	2.5+	76+	4.407
45+	M	3.5+	76+	15.448
45+	F	3.5+	76+	4.827
45+	M	4.5+	76+	2.947
45+	F	4.5+	76+	1.141
45+	M	0.0+	73+	2.393
45+	F	0.0+	73+	1.110
45+	M	2.5+	73+	10.806
45+	F	2.5+	73+	5.499
45+	M	3.5+	73+	16.735
45+	F	3.5+	73+	6.616
45+	M	4.5+	73+	4.094
45+	F	4.5+	73+	2.100
45+	M	0.0+	69+	2.044
45+	F	0.0+	69+	1.026
45+	M	2.5+	69+	9.585
45+	F	2.5+	69+	4.405
45+	M	3.5+	69+	11.977
45+	F	3.5+	69+	3.978
45+	M	4.5+	69+	2.611
45+	F	4.5+	69+	1.320
45+	M	0.0+	65+	0.506
45+	F	0.0+	65+	0.273
45+	M	2.5+	65+	4.069
45+	F	2.5+	65+	2.225
45+	M	3.5+	65+	5.422
45+	F	3.5+	65+	2.674
45+	M	4.5+	65+	0.633
45+	F	4.5+	65+	0.460
45+	M	0.0+	0+	0.171
45+	F	0.0+	0+	0.055
45+	M	2.5+	0+	1.443
45+	F	2.5+	0+	0.410
45+	M	3.5+	0+	1.952
45+	F	3.5+	0+	0.425
45+	M	4.5+	0+	0.107
45+	F	4.5+	0+	0.071
55+	M	0.0+	76+	1.887
55+	F	0.0+	76+	0.545
55+	M	2.5+	76+	7.075
55+	F	2.5+	76+	2.761
55+	M	3.5+	76+	15.189
55+	F	3.5+	76+	4.657
55+	M	4.5+	76+	2.299
55+	F	4.5+	76+	0.461
55+	M	0.0+	73+	2.087
55+	F	0.0+	73+	0.640
55+	M	2.5+	73+	9.499
55+	F	2.5+	73+	3.981
55+	M	3.5+	73+	14.095
55+	F	3.5+	73+	5.070
55+	M	4.5+	73+	4.954
55+	F	4.5+	73+	1.242
55+	M	0.0+	69+	2.375
55+	F	0.0+	69+	0.714
55+	M	2.5+	69+	8.510
55+	F	2.5+	69+	2.921
55+	M	3.5+	69+	14.784
55+	F	3.5+	69+	4.237
55+	M	4.5+	69+	3.872
55+	F	4.5+	69+	0.866
55+	M	0.0+	65+	1.396
55+	F	0.0+	65+	0.395
55+	M	2.5+	65+	4.237
55+	F	2.5+	65+	1.513
55+	M	3.5+	65+	5.294
55+	F	3.5+	65+	1.954
55+	M	4.5+	65+	1.145
55+	F	4.5+	65+	0.323
55+	M	0.0+	0+	0.490
55+	F	0.0+	0+	0.162
55+	M	2.5+	0+	0.936
55+	F	2.5+	0+	0.338
55+	M	3.5+	0+	1.134
55+	F	3.5+	0+	0.337
55+	M	4.5+	0+	0.245
55+	F	4.5+	0+	0.121

TABLE 9. TWO-DIMENSIONAL MARGINS (Annual Billions of VMT, 1977 NPTS)

	SEX	
AGE	MALE	FEMALE
0-24	106.825	60.403
25-34	158.835	70.542
35-44	114.798	51.354
45-54	105.224	44.073
55 and over	101.674	33.034

Age of Driver vs. Sex of Driver

	Weight of Vehicle (1000 lbs.)			
AGE	0-2.5	2.5-3.5	3.5-4.5	over 4.5
0-24	20.183	64.327	72.362	10.356
25-34	27.500	85.921	101.562	14.394
35-44	18.582	53.676	78.068	15.826
45-54	11.220	52.715	69.847	15.484
55 and over	10.690	41.780	66.720	15.518

Age of Driver vs. Weight of Vehicle

	YEAR OF MODEL				
	76-78	73-75	69-72	65-68	64 or older
AGE					
0-24	35.623	53.654	47.037	22.456	8.457
25-34	58.871	76.084	61.905	24.040	8.478
35-44	44.468	50.687	46.763	18.277	5.957
45-54	42.309	49.354	36.938	16.062	4.634
55 or older	35.058	41.570	38.092	16.256	3.733

Age of Driver vs. Year of Model

TABLE 10. TWO-DIMENSIONAL MARGINS (CONT.)
(Annual Billions of VMT, 1977 NPTS)

Weight of Vehicle (1000 lbs.)

SEX	0.25	2.5-3.5	3.5-4.5	over 4.5
MALE	60.257	202.554	275.087	49.458
FEMALE	27.918	95.896	113.471	22.121

Sex of Driver vs. Weight of Vehicle

	YEAR OF MODEL				
SEX	76-78	73-75	69-72	65-68	64 or older
MALE	153.812	181.826	157.820	68.872	25.026
FEMALE	62.518	89.523	72.915	28.218	6.232

Sex of Driver vs. Year of Model

	YEAR OF MODEL				
WEIGHT OF VEHICLE	76-78	73-75	69-72	65-68	64 or older
0-2.5	27.127	29.547	21.338	6.479	.3.583
2.5-2.5	67.943	91.043	87.777	39.715	11.972
3.5-4.5	105.250	123.443	102.846	42.898	14.122
over 4.5	16.010	27.316	18.774	7.997	1.481

Weight of Vehicle vs. Year of Model

TABLE 11. CELL ESTIMATES FROM A LOG LINEAR MODEL BASED ON 2-DIMENSIONAL MARGINS

DRIVER AGE	DRIVER SEX	VEHICLE WGT.	VEHICLE YR.	VMT BILLION
0+	M	0.0+	76+	3.435
0+	F	0.0+	76+	1.871
0+	M	2.5+	76+	7.745
0+	F	2.5+	76+	4.363
0+	M	3.5+	76+	10.892
0+	F	3.5+	76+	5.423
0+	M	4.5+	76+	1.227
0+	F	4.5+	76+	0.668
0+	M	0.0+	73+	4.181
0+	F	0.0+	73+	2.721
0+	M	2.5+	73+	11.666
0+	F	2.5+	73+	7.850
0+	M	3.5+	73+	14.571
0+	F	3.5+	73+	8.666
0+	M	4.5+	73+	2.423
0+	F	4.5+	73+	1.575
0+	M	0.0+	69+	3.162
0+	F	0.0+	69+	1.926
0+	M	2.5+	69+	11.868
0+	F	2.5+	69+	7.476
0+	M	3.5+	69+	12.721
0+	F	3.5+	69+	7.982
0+	M	4.5+	69+	1.741
0+	F	4.5+	69+	1.060
0+	M	0.0+	65+	1.145
0+	F	0.0+	65+	0.613
0+	M	2.5+	65+	6.392
0+	F	2.5+	65+	3.542
0+	M	3.5+	65+	6.313
0+	F	3.5+	65+	3.091
0+	M	4.5+	65+	0.886
0+	F	4.5+	65+	0.474
0+	M	0.0+	0+	0.456
0+	F	0.0+	0+	0.272
0+	M	2.5+	0+	2.519
0+	F	2.5+	0+	0.846
0+	M	3.5+	0+	2.793
0+	F	3.5+	0+	0.810
0+	M	4.5+	0+	0.229
0+	F	4.5+	0+	0.073
25+	M	0.0+	76+	5.988
25+	F	0.0+	76+	2.542
25+	M	2.5+	76+	13.695
25+	F	2.5+	76+	6.014
25+	M	3.5+	76+	19.788
25+	F	3.5+	76+	7.681
25+	M	4.5+	76+	2.221
25+	F	4.5+	76+	0.942
25+	M	0.0+	73+	6.316
25+	F	0.0+	73+	3.204
25+	M	2.5+	73+	17.879
25+	F	2.5+	73+	9.378
25+	M	3.5+	73+	22.943
25+	F	3.5+	73+	10.637
25+	M	4.5+	73+	3.801
25+	F	4.5+	73+	1.926
25+	M	0.0+	69+	4.819
25+	F	0.0+	69+	2.098
25+	M	2.5+	69+	16.826
25+	F	2.5+	69+	8.262
25+	M	3.5+	69+	18.530
25+	F	3.5+	69+	8.042
25+	M	4.5+	69+	2.527
25+	F	4.5+	69+	1.199
25+	M	0.0+	65+	1.292
25+	F	0.0+	65+	0.539
25+	M	2.5+	65+	7.317
25+	F	2.5+	65+	3.160
25+	M	3.5+	65+	7.425
25+	F	3.5+	65+	2.834
25+	M	4.5+	65+	1.036
25+	F	4.5+	65+	0.433
25+	M	0.0+	0+	0.881
25+	F	0.0+	0+	0.218
25+	M	2.5+	0+	2.899
25+	F	2.5+	0+	0.690
25+	M	3.5+	0+	3.003
25+	F	3.5+	0+	0.679
25+	M	4.5+	0+	0.345
25+	F	4.5+	0+	0.061
35+	M	0.0+	76+	4.209
35+	F	0.0+	76+	1.813
35+	M	2.5+	76+	8.854
35+	F	2.5+	76+	3.945
35+	M	3.5+	76+	15.780
35+	F	3.5+	76+	6.215
35+	M	4.5+	76+	2.553
35+	F	4.5+	76+	1.099
35+	M	0.0+	73+	3.883
35+	F	0.0+	73+	1.998
35+	M	2.5+	73+	10.119
35+	F	2.5+	73+	5.381
35+	M	3.5+	73+	16.002
35+	F	3.5+	73+	7.527
35+	M	4.5+	73+	3.821
35+	F	4.5+	73+	1.965
35+	M	0.0+	69+	3.125
35+	F	0.0+	69+	1.505
35+	M	2.5+	69+	10.942
35+	F	2.5+	69+	5.452
35+	M	3.5+	69+	14.864
35+	F	3.5+	69+	6.546
35+	M	4.5+	69+	2.922
35+	F	4.5+	69+	1.407
35+	M	0.0+	65+	0.921
35+	F	0.0+	65+	0.390
35+	M	2.5+	65+	4.801
35+	F	2.5+	65+	2.104
35+	M	3.5+	65+	6.009
35+	F	3.5+	65+	2.327
35+	M	4.5+	65+	1.211
35+	F	4.5+	65+	0.513
35+	M	0.0+	0+	0.598
35+	F	0.0+	0+	0.148
35+	M	2.5+	0+	1.658
35+	F	2.5+	0+	0.430
35+	M	3.5+	0+	2.275
35+	F	3.5+	0+	0.522
35+	M	4.5+	0+	0.288
35+	F	4.5+	0+	0.067
45+	M	0.0+	76+	2.753
45+	F	0.0+	76+	1.195
45+	M	2.5+	76+	9.600
45+	F	2.5+	76+	3.987
45+	M	3.5+	76+	15.402
45+	F	3.5+	76+	5.654
45+	M	4.5+	76+	2.718
45+	F	4.5+	76+	1.091
45+	M	0.0+	73+	2.577
45+	F	0.0+	73+	1.236
45+	M	2.5+	73+	11.121
45+	F	2.5+	73+	5.517
45+	M	3.5+	73+	15.847
45+	F	3.5+	73+	6.948
45+	M	4.5+	73+	4.128
45+	F	4.5+	73+	1.979
45+	M	0.0+	69+	1.666
45+	F	0.0+	69+	0.748
45+	M	2.5+	69+	9.673
45+	F	2.5+	69+	4.492
45+	M	3.5+	69+	11.829
45+	F	3.5+	69+	4.855
45+	M	4.5+	69+	2.537
45+	F	4.5+	69+	1.138
45+	M	0.0+	65+	0.539
45+	F	0.0+	65+	0.213
45+	M	2.5+	65+	4.667
45+	F	2.5+	65+	1.902
45+	M	3.5+	65+	5.247
45+	F	3.5+	65+	1.894
45+	M	4.5+	65+	1.154
45+	F	4.5+	65+	0.455
45+	M	0.0+	0+	0.310
45+	F	0.0+	0+	0.073
45+	M	2.5+	0+	1.447
45+	F	2.5+	0+	0.350
45+	M	3.5+	0+	1.788
45+	F	3.5+	0+	0.382
45+	M	4.5+	0+	0.229
45+	F	4.5+	0+	0.054
55+	M	0.0+	76+	2.600
55+	F	0.0+	76+	0.811
55+	M	2.5+	76+	7.364
55+	F	2.5+	76+	2.377
55+	M	3.5+	76+	14.328
55+	F	3.5+	76+	4.087
55+	M	4.5+	76+	2.661
55+	F	4.5+	76+	0.830
55+	M	0.0+	73+	2.500
55+	F	0.0+	73+	0.932
55+	M	2.5+	73+	8.763
55+	F	2.5+	73+	3.378
55+	M	3.5+	73+	15.147
55+	F	3.5+	73+	5.159
55+	M	4.5+	73+	4.151
55+	F	4.5+	73+	1.546
55+	M	0.0+	69+	1.992
55+	F	0.0+	69+	0.695
55+	M	2.5+	69+	9.195
55+	F	2.5+	69+	3.395
55+	M	3.5+	69+	13.933
55+	F	3.5+	69+	4.444
55+	M	4.5+	69+	3.145
55+	F	4.5+	69+	1.097
55+	M	0.0+	65+	0.632
55+	F	0.0+	65+	0.194
55+	M	2.5+	65+	4.434
55+	F	2.5+	65+	1.407
55+	M	3.5+	65+	6.058
55+	F	3.5+	65+	1.699
55+	M	4.5+	65+	1.402
55+	F	4.5+	65+	0.437
55+	M	0.0+	0+	0.283
55+	F	0.0+	0+	0.051
55+	M	2.5+	0+	1.071
55+	F	2.5+	0+	0.201
55+	M	3.5+	0+	1.664
55+	F	3.5+	0+	0.266
55+	M	4.5+	0+	0.217
55+	F	4.5+	0+	0.039

TABLE 12. ONE-DIMENSIONAL MARGINS (Annual Billions of VMT, 1977 NPTS)

0-24	25-34	35-44	45-54	55 and over
167.227	229.378	166.152	149.297	134.709

Age of Driver

MALE	FEMALE
587.356	259.406

Sex of Driver

0-2.5	2.5-3.5	3.5-4.5	over 4.5
88.175	298.450	388.558	71.579

Weight of Vehicle (1000 lbs.)

76-78	73-75	69-72	65-68	64 or older
216.330	271.349	230.735	97.089	36.259

Year of Model

TABLE 13. CELL ESTIMATES FROM LOG LINEAR MODEL BASED ON 1-DIMENSIONAL MARGINS

DRIVER AGE	DRIVER SEX	VEHICLE WGT.	VEHICLE YR.	VMT BILLION
0+	M	0.0+	76+	3.086
0+	F	0.0+	76+	1.363
0+	M	2.5+	76+	10.445
0+	F	2.5+	76+	4.613
0+	M	3.5+	76+	13.599
0+	F	3.5+	76+	6.006
0+	M	4.5+	76+	2.505
0+	F	4.5+	76+	1.106
0+	M	0.0+	73+	3.871
0+	F	0.0+	73+	1.710
0+	M	2.5+	73+	13.102
0+	F	2.5+	73+	5.786
0+	M	3.5+	73+	17.057
0+	F	3.5+	73+	7.533
0+	M	4.5+	73+	3.142
0+	F	4.5+	73+	1.388
0+	M	0.0+	69+	3.291
0+	F	0.0+	69+	1.454
0+	M	2.5+	69+	11.141
0+	F	2.5+	69+	4.920
0+	M	3.5+	69+	14.504
0+	F	3.5+	69+	6.406
0+	M	4.5+	69+	2.672
0+	F	4.5+	69+	1.180
0+	M	0.0+	65+	1.385
0+	F	0.0+	65+	0.612
0+	M	2.5+	65+	4.688
0+	F	2.5+	65+	2.070
0+	M	3.5+	65+	6.103
0+	F	3.5+	65+	2.695
0+	M	4.5+	65+	1.124
0+	F	4.5+	65+	0.497
0+	M	0.0+	0+	0.446
0+	F	0.0+	0+	0.197
0+	M	2.5+	0+	1.509
0+	F	2.5+	0+	0.667
0+	M	3.5+	0+	1.965
0+	F	3.5+	0+	0.868
0+	M	4.5+	0+	0.362
0+	F	4.5+	0+	0.160

DRIVER AGE	DRIVER SEX	VEHICLE WGT.	VEHICLE YR.	VMT BILLION
25+	M	0.0+	76+	4.233
25+	F	0.0+	76+	1.869
25+	M	2.5+	76+	14.327
25+	F	2.5+	76+	6.328
25+	M	3.5+	76+	18.653
25+	F	3.5+	76+	8.238
25+	M	4.5+	76+	3.436
25+	F	4.5+	76+	1.518
25+	M	0.0+	73+	5.309
25+	F	0.0+	73+	2.345
25+	M	2.5+	73+	17.971
25+	F	2.5+	73+	7.937
25+	M	3.5+	73+	23.397
25+	F	3.5+	73+	10.333
25+	M	4.5+	73+	4.310
25+	F	4.5+	73+	1.904
25+	M	0.0+	69+	4.515
25+	F	0.0+	69+	1.994
25+	M	2.5+	69+	15.281
25+	F	2.5+	69+	6.749
25+	M	3.5+	69+	19.895
25+	F	3.5+	69+	8.787
25+	M	4.5+	69+	3.665
25+	F	4.5+	69+	1.619
25+	M	0.0+	65+	1.903
25+	F	0.0+	65+	0.839
25+	M	2.5+	65+	6.430
25+	F	2.5+	65+	2.840
25+	M	3.5+	65+	8.371
25+	F	3.5+	65+	3.697
25+	M	4.5+	65+	1.542
25+	F	4.5+	65+	0.681
25+	M	0.0+	0+	0.612
25+	F	0.0+	0+	0.270
25+	M	2.5+	0+	2.074
25+	F	2.5+	0+	0.914
25+	M	3.5+	0+	2.695
25+	F	3.5+	0+	1.190
25+	M	4.5+	0+	0.497
25+	F	4.5+	0+	0.219

DRIVER AGE	DRIVER SEX	VEHICLE WGT.	VEHICLE YR.	VMT BILLION
35+	M	0.0+	76+	3.066
35+	F	0.0+	76+	1.354
35+	M	2.5+	76+	10.378
35+	F	2.5+	76+	4.583
35+	M	3.5+	76+	13.511
35+	F	3.5+	76+	5.967
35+	M	4.5+	76+	2.489
35+	F	4.5+	76+	1.099
35+	M	0.0+	73+	3.846
35+	F	0.0+	73+	1.699
35+	M	2.5+	73+	13.017
35+	F	2.5+	73+	5.749
35+	M	3.5+	73+	16.948
35+	F	3.5+	73+	7.485
35+	M	4.5+	73+	3.122
35+	F	4.5+	73+	1.379
35+	M	0.0+	69+	3.270
35+	F	0.0+	69+	1.444
35+	M	2.5+	69+	11.069
35+	F	2.5+	69+	4.889
35+	M	3.5+	69+	14.411
35+	F	3.5+	69+	6.365
35+	M	4.5+	69+	2.655
35+	F	4.5+	69+	1.172
35+	M	0.0+	65+	1.376
35+	F	0.0+	65+	0.608
35+	M	2.5+	65+	4.658
35+	F	2.5+	65+	2.057
35+	M	3.5+	65+	6.064
35+	F	3.5+	65+	2.678
35+	M	4.5+	65+	1.117
35+	F	4.5+	65+	0.493
35+	M	0.0+	0+	0.443
35+	F	0.0+	0+	0.196
35+	M	2.5+	0+	1.500
35+	F	2.5+	0+	0.662
35+	M	3.5+	0+	1.952
35+	F	3.5+	0+	0.862
35+	M	4.5+	0+	0.360
35+	F	4.5+	0+	0.159

DRIVER AGE	DRIVER SEX	VEHICLE WGT.	VEHICLE YR.	VMT BILLION
45+	M	0.0+	76+	2.755
45+	F	0.0+	76+	1.217
45+	M	2.5+	76+	9.325
45+	F	2.5+	76+	4.118
45+	M	3.5+	76+	12.141
45+	F	3.5+	76+	5.362
45+	M	4.5+	76+	2.236
45+	F	4.5+	76+	0.988
45+	M	0.0+	73+	3.456
45+	F	0.0+	73+	1.528
45+	M	2.5+	73+	11.697
45+	F	2.5+	73+	5.166
45+	M	3.5+	73+	15.228
45+	F	3.5+	73+	6.728
45+	M	4.5+	73+	2.805
45+	F	4.5+	73+	1.239
45+	M	0.0+	69+	2.939
45+	F	0.0+	69+	1.298
45+	M	2.5+	69+	9.946
45+	F	2.5+	69+	4.393
45+	M	3.5+	69+	12.949
45+	F	3.5+	69+	5.719
45+	M	4.5+	69+	2.385
45+	F	4.5+	69+	1.054
45+	M	0.0+	65+	1.236
45+	F	0.0+	65+	0.546
45+	M	2.5+	65+	4.185
45+	F	2.5+	65+	1.848
45+	M	3.5+	65+	5.449
45+	F	3.5+	65+	2.406
45+	M	4.5+	65+	1.004
45+	F	4.5+	65+	0.443
45+	M	0.0+	0+	0.398
45+	F	0.0+	0+	0.176
45+	M	2.5+	0+	1.347
45+	F	2.5+	0+	0.595
45+	M	3.5+	0+	1.754
45+	F	3.5+	0+	0.775
45+	M	4.5+	0+	0.323
45+	F	4.5+	0+	0.143

DRIVER AGE	DRIVER SEX	VEHICLE WGT.	VEHICLE YR.	VMT BILLION
55+	M	0.0+	76+	2.486
55+	F	0.0+	76+	1.098
55+	M	2.5+	76+	8.414
55+	F	2.5+	76+	3.716
55+	M	3.5+	76+	10.954
55+	F	3.5+	76+	4.838
55+	M	4.5+	76+	2.018
55+	F	4.5+	76+	0.891
55+	M	0.0+	73+	3.118
55+	F	0.0+	73+	1.377
55+	M	2.5+	73+	10.554
55+	F	2.5+	73+	4.661
55+	M	3.5+	73+	13.749
55+	F	3.5+	73+	6.068
55+	M	4.5+	73+	2.531
55+	F	4.5+	73+	1.118
55+	M	0.0+	69+	2.651
55+	F	0.0+	69+	1.171
55+	M	2.5+	69+	8.974
55+	F	2.5+	69+	3.961
55+	M	3.5+	69+	11.684
55+	F	3.5+	69+	5.159
55+	M	4.5+	69+	2.152
55+	F	4.5+	69+	0.951
55+	M	0.0+	65+	1.116
55+	F	0.0+	65+	0.493
55+	M	2.5+	65+	3.776
55+	F	2.5+	65+	1.668
55+	M	3.5+	65+	4.916
55+	F	3.5+	65+	2.171
55+	M	4.5+	65+	0.906
55+	F	4.5+	65+	0.400
55+	M	0.0+	0+	0.359
55+	F	0.0+	0+	0.159
55+	M	2.5+	0+	1.216
55+	F	2.5+	0+	0.537
55+	M	3.5+	0+	1.583
55+	F	3.5+	0+	0.699
55+	M	4.5+	0+	0.292
55+	F	4.5+	0+	0.129

The standard statistics for the three models are shown in Table 14. As noted earlier, the chi-square and likelihood-ratio statistics are not valid; in fact, they are derived from VMT cell entries and are thus grossly distorted (and inflated) by weights and expansion factors (See Section 3.2.4 and Appendix A for a suggested correction factor estimates). Nevertheless, the statistics are a crude measure of relative goodness of fit, so that one may observe that the 3-level model appears to be the best fit;* the 1-level model the worst fit of the three models. The degrees of freedom, on the other hand, are valid since they depend only on the total number of cells and the number of parameters utilized in the model. If one assumes that the inflation factor in the VMT data is approximately constant across cells, one can observe that the ratio of chi-square (2.963) and of D.O.F. (2.833) for models 2 and 1 are almost equal. This can mean that chi-square is measuring noise. This suggests that the better fit for model 1 is not statistically significant. The same crude test tells us with somewhat more confidence that model 3 is probably not as good a fit as models 1 or 2. Of course, these observations are highly intuitive and qualitative. Rigorous goodness-of-fit tests cannot be applied unless successful scaling adjustments can be made to the VMT data, or to the chi-square statistics.

Another observation that can be made from these tests has to do with the ability of log-linear models to detect data instabilities. In particular, if one divides the given data set of 200 cells into two sets of 100 cells each, corresponding to male drivers and female drivers respectively, one can pair the cells naturally i.e., for each "female" cell, there is exactly one "male" cell having the same levels for all variables except sex. Inspection of Table 3 reveals that for 91 of these cell-pairs the VMT for the "male" cell is much greater than the VMT for the corresponding "female" cell. In 9 of the cell-pairs, however, the "female" VMT is larger or close to equal to the "male" VMT. (This can be verified easily since Table 3 is arranged so that corresponding cells are adjacent.) If one now looks at the cell estimates for the three models, it is observed that, in each case, all 9 cell-pairs are reversed, so that the "male" cell dominates for all 100 cell-pairs. This, in turn prompts an investigation of the marginal tables which reveals that for all margins (3, 2, and 1-dimensional) the "male" cell has a

*I.e the closest fit as measured by chi-square. It could not be otherwise since the all three factor model contains the others.

TABLE 14. CHI-SQUARE, G^2 AND DEGREES OF FREEDOM FOR THREE HOMOGENEOUS HIERARCHICAL LOG LINEAR MODELS

	CHI-SQUARE	G^2	D.O.F.
MODEL 1 (3-variable effects)	$.707 \times 10^{10}$	$.715 \times 10^{10}$	48
MODEL 2 (2-variable effects)	2.095×10^{10}	2.050×10^{10}	136
MODEL 3 (1-variable effects)	5.295×10^{10}	5.227×10^{10}	187

substantially larger VMT than the corresponding "female" cell for all cell pairs. Thus, it is not surprising that the (unsaturated) log linear models, which are estimated from marginal information alone, reversed the male/female VMT ratio in the 9 anomolous cell-pairs. One can now zero in on the more disaggregate data to look for evidence of something wrong in the VMT values for these cells. In fact, preliminary evidence that there may be some thin sampling problems in these cells can be obtained from Table 3, by observing that each of the cells correspond to the lightest (0.0+) and heaviest (4.5+) auto classes which, according to Table 12, comprise approximately 10.4 percent and 8.5 percent of the population, respectively.

In summary, we have prepared a modest-sized multivariate data set, for test purposes, from the 1977 NPTS tapes. We have formally fit a sample set of unsaturated hierarchical log linear models to the data in order to demonstrate and test the capability.

TABLE 15. TWO-DIMENSIONAL MARGINS (Percentage Distributions of VMT)

	AGE		
SEX	24	25-54	55
Male	7.1	39.8	12.8
Female	6.6	28.0	5.7

	AGE		
TIME	24	25-54	55
Day	9.3	52.1	15.5
Night	4.4	15.7	3.0

	AGE		
PLACE	24	25-54	55
Urban	6.6	36.3	10.0
Rural	7.1	31.5	8.5

	AGE		
YEAR OF MODEL	24	25-54	55
72-74	5.0	21.4	6.9
69-71	3.6	22.0	4.3
68	5.1	24.5	7.3

SOURCE: White, S. B., C. A. Clayton, L. D. Bressler and J. R. Stewart, "Improved Exposure Measures" Final Report, Research Triangle Institute, Contract No. DOT-HS-022-2-418, September, 1975.

TABLE 15. TWO-DIMENSIONAL MARGINS (CONT.)
(Percentage Distribution of VMT)

TIME	SEX	
	MALE	FEMALE
Day	44.0	32.9
Night	15.7	7.5

PLACE	SEX	
	MALE	FEMALE
Urban	29.4	23.5
Rural	30.2	16.9

YEAR OF MODEL	SEX	
	MALE	FEMALE
72-74	20.3	12.9
69-71	15.9	14.0
68	23.4	13.4

YEAR OF MODEL	TIME	
	DAY	NIGHT
72-74	25.4	7.8
69-71	22.4	7.5
68	29.1	7.8

YEAR OF MODEL	PLACE	
	URBAN	RURAL
72-74	16.9	16.4
69-71	17.6	12.3
68	18.5	18.4

TIME	PLACE	
	URBAN	RURAL
Day	39.7	37.1
Night	13.2	10.0

SOURCE: White, Clayton, Bressler and Stewart

TABLE 16. 5-WAY OUTPUT FROM IPF (Percentage Distributions of VMT)

YEAR OF MODEL	TIME	PLACE	MALE < 25	MALE 25-55	MALE > 55	FEMALE < 24	FEMALE 25-55	FEMALE > 55
72-74	D	U	0.65	4.30	1.90	0.84	3.84	1.06
		R	0.98	5.07	2.01	0.88	3.11	0.77
	N	U	0.47	1.85	0.50	0.35	0.95	0.16
		R	0.54	1.67	0.41	0.28	0.59	0.09
69-71	D	U	0.45	4.33	1.20	0.79	5.21	0.90
		R	0.51	3.79	0.94	0.61	3.13	0.49
	N	U	0.36	2.06	0.35	0.36	1.42	0.15
		R	0.31	1.38	0.21	0.21	0.65	0.06
< 68	D	U	0.72	5.28	2.11	0.84	4.23	1.05
		R	1.11	6.29	2.27	0.89	3.47	0.78
	N	U	0.45	1.96	0.48	0.30	0.90	0.14
		R	0.53	1.79	0.39	0.24	0.57	0.08

TABLE 17. COMPARISON OF TABLES FOR AGE, TIME AND YEAR OF MODEL
(Percentage Distributions of VMT)

YEAR OF MODEL	AGE	≤ 24		25-54		≥ 55	
	TIME	Day	Night	Day	Night	Day	Night
72-74		3.6	1.3	16.1	5.3	5.6	1.2
69-71		2.4	1.2	16.6	5.4	3.4	0.9
< 68		3.3	1.8	19.4	5.1	6.4	0.9

3-WAY TABLE - ORIGINAL DATA

SOURCE: White, Clayton, Bressler and Stewart

YEAR OF MODEL	AGE	≤ 24		25-54		≥ 55	
	TIME	Day	Night	Day	Night	Day	Night
72-74		3.4	1.6	16.3	5.1	5.7	1.2
69-71		2.4	1.2	16.5	5.5	3.5	0.8
< 68		3.6	1.5	19.3	5.2	6.2	1.1

3-DIMENSIONAL MARGIN - 5-WAY IPF OUTPUT TABLE

YEAR OF MODEL	AGE	≤ 24		25-54		≥ 55	
	TIME	Day	Night	Day	Night	Day	Night
72-74		3.5	1.1	17.3	5.2	4.7	1.4
69-71		3.2	0.9	15.6	4.7	4.3	1.3
< 68		3.9	1.2	19.2	5.8	5.2	1.6

INDEPENDENT 3-WAY TABLE

TABLE 18. COMPARISON OF TABLES FOR SEX, AGE AND YEAR OF MODEL
(Percentage Distributions of VMT)

YEAR OF MODEL	SEX / AGE	MALE			FEMALE		
		<24	25-54	>55	<24	25-54	>55
72-74		2.6	12.7	5.1	2.4	8.7	1.8
69-71		1.7	11.7	2.6	1.9	10.3	1.7
68		2.8	15.5	5.1	2.3	9.0	2.2

3-WAY TABLE - ORIGINAL DATA

SOURCE: White, Clayton, Bressler and Stewart

YEAR OF MODEL	SEX / AGE	MALE			FEMALE		
		<24	25-54	>55	<24	25-54	>55
72-74		2.6	12.9	4.8	2.4	8.5	2.1
69-71		1.6	11.6	2.7	2.0	10.4	1.6
68		2.8	15.3	5.3	2.3	9.2	2.0

3-DIMENSIONAL MARGIN FROM 5-WAY IPF OUTPUT

YEAR OF MODEL	SEX / AGE	MALE			FEMALE		
		<24	25-54	>55	<24	25-54	>55
72-74		2.7	13.4	3.7	1.8	9.1	2.5
69-71		2.4	12.1	3.3	1.7	8.2	2.2
68		3.0	14.9	4.1	2.0	10.1	2.8

INDEPENDENT 3-WAY TABLE

TABLE 19. COMPARISONS OF 3-WAY TABLES FOR SEX, TIME AND YEAR OF MODEL (Percentage Distributions of VMT)

YEAR OF MODEL	SEX / TIME	MALE		FEMALE	
		Day	Night	Day	Night
72-74		15.0	5.3	10.4	2.5
69-71		11.3	4.6	11.1	2.9
68		17.7	5.7	11.3	2.1

3-WAY TABLE - ORIGINAL DATA

SOURCE: White, Clayton, Bressler and Stewart

YEAR OF MODEL	SEX / TIME	MALE		FEMALE	
		Day	Night	Day	Night
72-74		14.9	5.4	10.5	2.4
69-71		11.2	4.7	11.1	2.9
68		17.8	5.6	11.3	2.2

3-DIMENSIONAL MARGIN FROM 5-WAY IPF OUTPUT

YEAR OF MODEL	SEX / TIME	MALE		FEMALE	
		Day	Night	Day	Night
72-74		15.2	4.6	10.3	3.1
69-71		13.7	4.1	9.3	2.8
68		16.9	5.1	11.5	3.4

INDEPENDENT 3-WAY TABLE

TABLE 20. U.S. REGISTERED DRIVERS BY AGE AND SEX, 1975, 1979, 1980 (1000's)

a) 1980 Actual

AGE	SEX MALE	FEMALE
0-24	16247	14285
25-34	18878	17417
35-44	12940	11888
45-54	10662	9504
55+	18463	15011

b) 1975 Actual

AGE	SEX MALE	FEMALE
0-24	15789	13533
25-34	15847	14215
35-44	11280	10090
45-54	11090	9493
55+	16552	11996

c) 1979 Actual

AGE	SEX MALE	FEMALE
0-24	16480	14381
25-34	18364	16819
35-44	12640	11557
45-54	10826	9561
55+	18222	14435

d) Independent 1980 Estimate
$X^2=84$

AGE	SEX MALE	FEMALE
0-24	16221	14311
25-34	19282	17012
34-44	13190	11638
45-54	10713	9452
55+	17783	15690

e) IPF 1980 Estimate
1975 Core $X^2=45$

AGE	SEX MALE	FEMALE
0-24	16060	14472
25-34	18680	17615
35-44	12838	11990
45-54	10613	9553
55+	18998	14476

f) IPF 1980 Estimate
1979 Core $X^2=2.3$

AGE	SEX MALE	FEMALE
0-24	16217	14315
25-34	18841	17456
35-44	12898	11930
45-54	10651	9515
55+	18583	14891

Source: Highway Statistics (U.S. DOT-FHWA)

TABLE 21. 1975 SATURATED LOG-LINEAR MODEL

U_0	9.45345684
$U_1(1)$	+.13652076
$U_1(2)$	+.16293741
$U_1(3)$	-.18189439
$U_1(4)$	-.21740229
$U_1(5)$	+.09983851
$U_2(1)$	+.08587358
$U_2(2)$	-.08587358
$U_{12}(11)$	-.00878238
$U_{12}(21)$	-.03153233
$U_{12}(31)$	-.02664953
$U_{12}(41)$	-.00812903
$U_{12}(51)$	+.07509327
$U_{12}(12)$	+.00878238
$U_{12}(22)$	+.03153233
$U_{12}(32)$	+.02664953
$U_{12}(42)$	+.00812903
$U_{12}(52)$	-.07509327

TABLE 22. 1980 SATURATED LOG-LINEAR MODEL

U_0	9.55989186
$U_1(1)$	+.07142259
$U_1(2)$	+.24558539
$U_1(3)$	-.13421016
$U_1(4)$	-.34293721
$U_1(5)$	+.16013939
$U_2(1)$	+.06160014
$U_2(2)$	-.06160014
$U_{12}(11)$	+.00274901
$U_{12}(21)$	-.02132489
$U_{12}(31)$	-.01920324
$U_{12}(41)$	-.00411349
$U_{12}(51)$	+.04189261
$U_{12}(12)$	-.00274901
$U_{12}(22)$	+.02132489
$U_{12}(32)$	+.01920324
$U_{12}(42)$	+.00411349
$U_{12}(52)$	-.04189261

TABLE 23. 1980/1975 SATURATED LOG-LINEAR MODEL

$U_0 = 9.55945323$

$U_1(1) = +.07257557$

$U_1(2) = +.24640532$

$U_1(3) = -.13345763$

$U_1(4) = -.34223593$

$U_1(5) = +.15671267$

$U_2(1) = +.06084445$

$U_2(2) = +.06084445$

$U_{12}(11) = -.00878235$

$U_{12}(21) = -.03153230$

$U_{12}(31) = -.02664955$

$U_{12}(41) = -.00812905$

$U_{12}(51) = +.07509325$

$U_{12}(12) = +.00878235$

$U_{12}(22) = +.03153230$

$U_{12}(32) = +.02664955$

$U_{12}(42) = +.00812905$

$U_{12}(52) = -.07509325$

TABLE 24. DIFFERENCES IN INTERACTIONS 1980 MODEL MINUS 1980/1975 MODEL

$U_1(1) = -.00115298$

$U_1(2) = -.00081993$

$U_1(3) = -.00075253$

$U_1(4) = -.00070128$

$U_1(5) = +.00342672$

$U_2(1) = +.00075569$

$U_2(2) = -.00075569$

$U_{12}(11) = +.01153136$

$U_{12}(21) = +.01020741$

$U_{12}(31) = +.00744631$

$U_{12}(41) = +.00401556$

$U_{12}(51) = -.03320064$

$U_{12}(12) = -.01153136$

$U_{12}(22) = -.01020741$

$U_{12}(32) = -.00744631$

$U_{12}(42) = -.00401556$

$U_{12}(52) = +.03320064$

4.2 SYNTHESIS OF MULTIVARIATE DATA SETS FROM LOWER DIMENSIONAL DATA (CASE 1 - DUMMY CORE)

This section describes the application of iterative proportional fitting (IPF) to the construction of a 5-dimensional data set, using a complete set of $\binom{5}{2}$ = ten 2-dimensional margins. The data are taken from a study[1] conducted in North Carolina in 1973, from which a table of VMT was prepared, classified by six variables: vehicle type, model year, driver sex, driver age, day/night and urban/rural. A 4-way table of the number of drivers, classified by the first four variables, was also prepared. The North Carolina (N.C.) report contains an extensive set of 2-way and 3-way tables of % distribution of VMT. The data in these tables formed the basis for the data synthesis example described below.

A set of ten 2-way tables of % distribution of VMT was constructed from the North Carolina report, one for each pair of variables from the following 5 variable set: model year, driver sex, driver age, day/night (time), and urban/rural (place). These tables are displayed in Table 15. The table was fitted to a five dimensional core of one's by IPF. The resulting 5-way table is shown as Table 16. For three subsets of three variables each (age, time, and model year; age, sex, and model year; and sex, time and model year, respectively), 3-way margins were computed from the IPF output table, and were compared with the corresponding 3-way tables taken from the N.C. report. In addition, for each case, the corresponding "independent" table (i.e., the table with no 2-factor or 3-factor effects) was computed. The three cases are shown in Tables 17, 18 and 19. As can be seen from these tables, in each case, the IPF procedure using ten 2-way margins with a dummy core estimated the 3-way tables quite well, certainly much better than the independent table. A further experiment was conducted. For each case, the three corresponding two-way margins were fitted by IPF to a 3-dimensional core of one's. In each case the output 3-way table was almost

---

[1] S.B. White, et al, "Improved Exposure Measurements" Final Report by Research Triangle Institute, under NHTSA Contract No. DOT-HS-022-2-418, September, 1975.

identical to the corresponding 3-dimensional margins computed from the synthesized 5-way table. We do not believe that the latter result is to be expected in general. It seems likely that there would be situations in which, for example, the 3-way margins computed from the 5-way table (synthesized from 10 2-way margins) would be significantly better estimates than the corresponding 3-way table synthesized from three 2-way margins.

It is of some interest to note that, if the ten 2-way margins were count data, the 5-way table derived from them by IPF would be the maximum likelihood cell estimates for an unsaturated hierarchical log linear model of any 5-way table of count data that yielded those margins.

The 3-way tables were the largest dimension of marginal tables published in the N.C. report. The complete data for the 6-way VMT table and the 4-way number-of-drivers table are in the appendix, but are largely illegible. If these data could be obtained from the authors, they could, after some manipulation in the computer, be used to conduct more extensive tests with IPF, including a direct comparison of the 5-way IPF table with the corresponding actual table.

4.3 SYNTHESIS OF MULTIVARIATE DATA SETS FROM LOWER DIMENSIONAL DATA (CASE 2 - ACTUAL CORE)

The final test of IPF involves the estimation of a two-dimensional table by fitting a fully classified core from a previous timeframe to the margins of a (more current) year of interest. The data were tables of registered drivers, classified by sex and age categories, compiled from 1975, 1979, and 1980 "Highway Statistics" (a FHWA publication). These tables are shown in Table 20, a, b, and c, respectively. The margins of the 1980 table were used to calculate the corresponding "independent" 2-way table (no 2-way interactions equivalent to fitting the margins to a core of 1's via IPF) shown in Table 20(d). The 1980 margins were fit via IPF to the 1975 Table 20(e) and the 1979, Table 20(f).

The main idea of the example was to see how well IPF performs with a core and, in particular, as an initial test of IPF's potential for updating a table with later year margins. (Conversely, of estimating a later year's tables given the later year margins and a core from an earlier year.) By comparing Table 20(a) with Tables d, e, and f, it is clear that fitting the 1975 core or the 1979 core gives better results than assuming independence. In the case of 1979, Table 20(f), the estimated table is very close to the 1980 table, in fact, the difference between the two tables is not significantly different from 0 at the 5% level. (The model estimates ten cells from 6 independent fixed margins, leaving 4 degrees of freedom. The value of X^2 at the 5% level for 4d.o.f. is 9.488.) In the case of the estimate based on a 1975 core, the difference between the 1980 and the estimated tables is completely determined by the different 2-way interactions in the 1975 table. Inspection of Tables 20(a) and e reveals that this effect is strongest on rows 1, 2, and especially 5, corresponding to the two youngest and the oldest age groups, respectively. These observations suggest changes in the sex composition of the corresponding age groups between 1975 and 1980.

The saturated log-linear model parameters were computed for the 1975 and 1980 matrices, and for the matrix derived from fitting the 1975 matrix to the 1980 margins (hereafter denoted the 1980/1975 matrix.) The parameters are shown in Tables 21, 22 and 23, respectively. Table 24 shows the difference in the parameters between the 1980 matrix and the 1980/1975 matrix. Inspection of this table reveals, as expected, that the second order interaction differences are dominant, and are strongest in the effect on the elements of rows 1 and 2 and especially 5. This is consistent with the differences observed in the matrices for 1980 and 1980/1975.

These calculations afford an opportunity to exemplify some of the properties of IPF described in section 3. For example, a comparison of Tables 21 and 23 shows that the 1975 second order interactions are preserved exactly (except for roundoff errors) by IPF. A comparison of Tables 22 and 23 shows that the 1980/1975 first-order interactions are slightly different. However,

the main result of section 3.3.3 implies that the 1980/1975 first-order interactions are completely independent of the first-order interactions of the 1975 (core) matrix. To illustrate this result by example, a matrix (Table 25) was constructed, having second order interactions exactly equal to those shown in Table 21 for the 1975 matrix, but with first-order and zero-order interactions equal to zero. When the matrix shown in Table 25 was fit to the 1980 margins, the outcome was, as predicted, exactly equal to the 1980/1975 matrix.

TABLE 25. CORE OF AGE-SEX DISTRIBUTIONS CORRESPONDING TO 1975 TABLE SECOND ORDER INTERACTIONS ONLY

AGE	SEX	
	MALE	FEMALE
0-24	.9912561	1.0088210
25-34	.9689596	1.0320347
35-44	.9737024	1.0270076
45-54	.9919039	1.0081621
55+	1.0779846	.9276570

APPENDIX A

FURTHER DETAILS ON APPLICATION OF LOG LINEAR MODELS TO CONTINUOUS DATA

The first question to be addressed in this Appendix is the effect of the classical (descrete) technique for estimating log linear models in the continuous case. Let V_{ijk} be the VMT for cell ijk (summed over all observations or records in the cell). Let X_{ijk} be the model estimate (say of $E(V_{ijk})$) for the cell.

The process of IPF applied to particular margins of V_{ijk} is the classical method of fitting log linear models to V_{ijk}. This process has two properties.

1. It maximizes

   $$\sum_{ijk} V_{ijk} \log_e (X_{ijk}/V_{ijk})$$

   (see Bishop et al Reference 1, p. 65.)

2. It makes the selected margins of X_{ijk} identical to those of V_{ijk}.

These two properties imply that the process minimizes

$$\sum_{ijk} V_{ijk} (X_{ijk}/V_{ijk} - \log_e (X_{ijk}/V_{ijk}))$$

(The latter quantity is related to the Poisson likelihood.)

This quantity may be reexpressed:

$$F(\underline{X}) = \sum_{ijk} V_{ijk}\, g\, (X_{ijk}/V_{ijk})$$

where

$$g(u) = u - \log_e u$$

then

$$g'(u) = 1 - 1/u$$

and

$$g''(u) = 1/u^2$$

It follows that F $(\underline{X})$ is convex in the model parameters X_{ijk}. It is the sum of terms each of which attains its minimum when $X_{ijk} = V_{ijk}$.

The function g (u) is sketched in Figure 2. g (u) attains its minimum when u=1. The criterion of fit used by this process is thus very much like the sum of the squares error criterion used in linear regression, having the following similar properties:

1. The criterion is a convex function of the model parameters.

2. It achieves its minimum when the model fits exactly.

The fact that the criterion is based on a sum of terms each dependent on only one data point is also similar in the two cases. In the present case similar to the case of least squares, each term is weighted by the magnitude of the data element. This is obvious from the presence of the V_{ijk} factor in the present case. The fact that large deviations are weighted much more heavily than small deviations is a familar property of least squares.

The chief difference is that g (u) goes to ∞ when u approaches zero. This prevents X_{ijk} from going to zero when V_{ijk} is positive and it also prevents negative values of X_{ijk}. These are necessary properties in the log linear modelling case.

The conclusion is that the fit criterion is very resaonable for general applications of log linear modelling independent of whether it gives a maximum likelihood estimate.

The second matter to be investigated in this Appendix is under what circumstances an actual maximim likelihood estimate (of cell means) can be obtained.

Let V_{ijk} denote the cell sum of VMT. Thus

$$V_{ijk} = \sum_{r=1}^{n} X_r$$

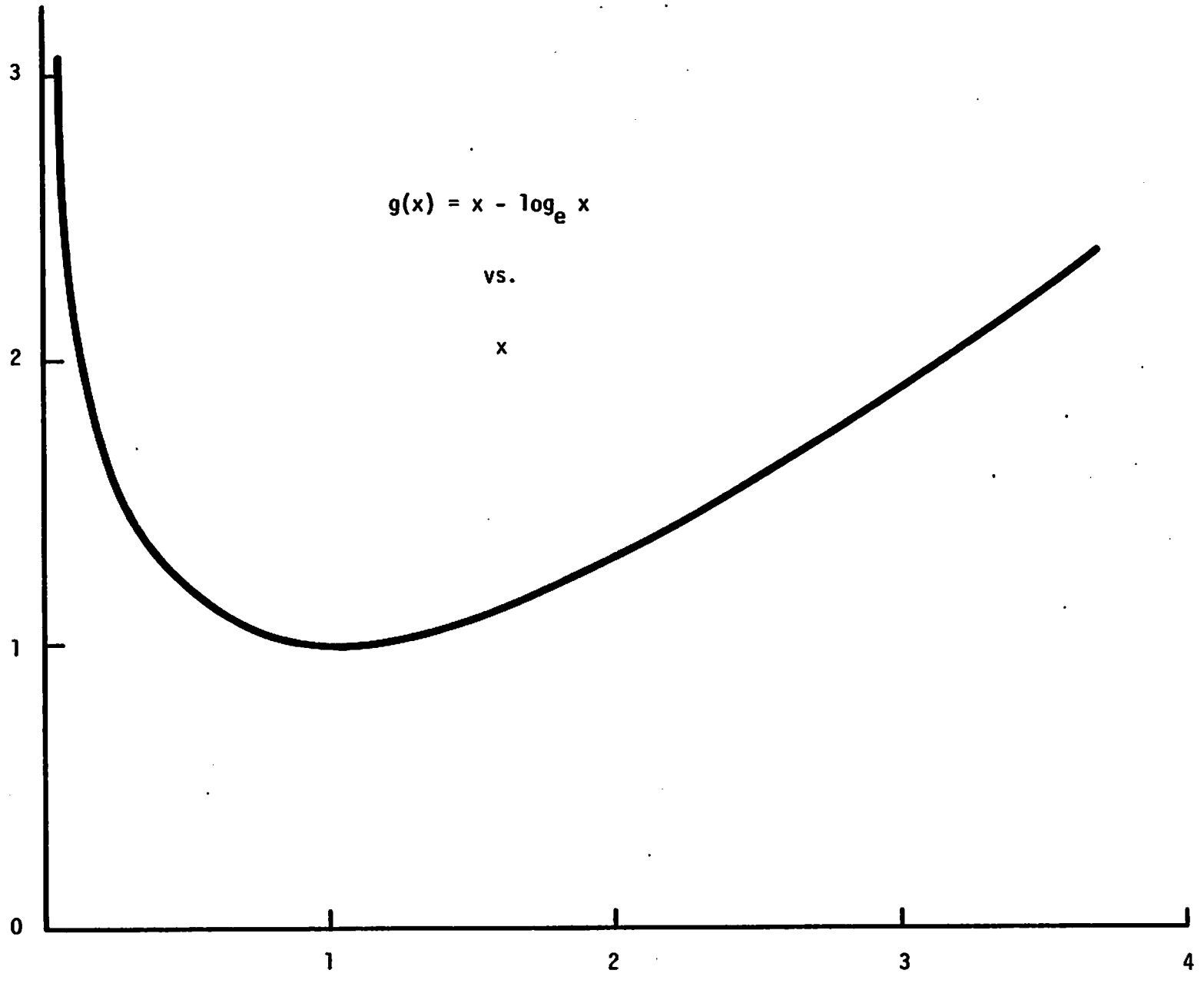

FIGURE 2. GRAPH OF A FUNCTION

where X_r denotes the weighted trip length of a given record in cell ijk. If n is fairly large and is Poisson distributed and independent of the X_r's (although both may be dependent on the choice of cell) then V_{ijk} will be approximately normally distributed. This will be assumed:

1. V_{ijk} is normally distributed with mean μ_{ijk} and variance σ^2_{ijk}.

If the distribution of V_{ijk} in each cell is characterized by two parameters then a likelihood function based on one parameter per cell cannot be constructed. In order to construct a likelihood function similar to the one maximized by the standard log linear model procedure it will be assumed that $\sigma^2_{ijk} = \alpha\ \mu_{ijk}$, i.e.,

2. The ratio of σ^2_{ijk} and μ_{ijk} does not depend on the cell (ijk).

On the assumption that V_{ijk} is normally distributed the log likelihood is easily written down:

$$L = -1/2 \sum_{ijk} (V_{ijk} - \mu_{ijk})^2/\sigma^2_{ijk} + \text{Constant}$$

where the constant does not depend on the data or the distribution parameters.

On the assumption that $\sigma^2_{ijk} = \alpha\mu_{ijk}$ this becomes

$$L = -1/(2\alpha) \sum_{ijk} (V_{ijk} - \mu_{ijk})^2/\mu_{ijk}$$

(dropping the constant as irrelevant to maximizing the likelihood).

The expression for $-2\alpha L$ is identical to the standard expression for the chi square statistic with respect to the model μ_{ijk}. If $\alpha = 1$ as it does in the case where V_{ijk} represents cell counts with a Poisson or multinomial distribution then what has been derived is the identity of -2L and chi square in the case of large samples. Since it is known that -2L has a chi square distribution in the case of large samples whatever the underlying distribution, it follows that the ordinary chi square statistic when divided by α will have a chi square distribution in the more general case.

Consequently if the number of observations in each cell is fairly large and if

$$\sigma^2_{ijk} = \alpha\mu_{ijk}$$

then

$$1/\alpha \sum_{ijk} (V_{ijk} - \mu_{ijk})^2/\mu_{ijk}$$

will have a chi square distribution and will serve as a statistical goodness of fit measure for the model μ_{ijk}.

Decisions regarding the appropriate degree of model complexity can be based on the chi square statistic. The relevant number of degrees of freedom can be calculated in the same manner as for standard discrete log linear models as discussed in Section 3.2.3.

The final matter to be considered in this Appendix is the question of whether the assumption that σ^2_{ijk} is proportional to μ_{ijk} is a good one and, if it is valid, how the constant of proportionality, α can be estimated.

Let

$$V_{ijk} = \sum_{r=1}^{n} X_r$$

Suppose that the X_r's are independent and identically distributed (within a given cell ijk) and that n is Poisson distributed and independent of X_r (although n and X_r both have distributions dependent on the cell ijk).

Let $\mu_{ijk} = E(V_{ijk})$

$$\sigma^2_{ijk} = E(V^2_{ijk}) - (E(V_{ijk}))^2$$

$$\bar{n} = E(n)$$

$$\overline{n^2} = E(n^2)$$

$$\sigma^2_x = E(X^2_r) - (E(X_r))^2$$

Then

$$E(V_{ijk}) = \sum_{n=0}^{\infty} P_r(n) \sum_{r=1}^{n} E(X_r)$$

$$= \sum_{n=0}^{\infty} P_r(n) \cdot \bar{n} \cdot E(X_r) = \bar{n}\, \mu_x$$

$$E(V_{ijk}^2) = \sum_{n=0}^{\infty} P_r(n)\, E\left(\sum_{r=1}^{n} X_r \sum_{s=1}^{n} X_s\right)$$

$$= \sum_{n=0}^{\infty} P_r(n)\left(n\, E(X_r^2) + (n^2-n)(E(X_r))^2\right)$$

$$= \bar{n}\, \sigma_x^2 + \overline{n^2}\, \mu_x^2$$

$$= \bar{n}\, \sigma_x^2 + \bar{n}\, \mu_x^2 + (\bar{n}\mu_x)^2$$

(since $\overline{n^2} = \bar{n} + (\bar{n})^2$ for a Poisson distribution)

Therefore

$$\mu_{ijk} = \bar{n}\mu_x$$

and

$$\sigma_{ijk}^2 = \bar{n}\,(\sigma_x^2 + \mu_x^2) = \bar{n}\, E(X_r^2)$$

Therefore

$$\alpha = \sigma_{ijk}^2/\mu_{ijk} = \sigma_x^2/\mu_x + \mu_x = E(x_r^2)/E(X_r)$$

For α to be constant across cells $E(x^2)/E(x)$ must be constant across cells. This is unlikely. However, it may happen that the greatest source of cell variance is the variance in n. In that case the assumption may not be too far off. In any case an estimate for α as if it were constant over cells is obtained as

$$\alpha = \left(\sum_{\text{cells}} \sum_{r} X_r^2\right) / \left(\sum_{\text{cells}} \sum_{r} X_r\right)$$

where the summations are over all records in <u>all</u> cells. This quantity is easily estimated before the rest of the log linear analysis takes place.

Clearly the assumption that $\sigma_{ijk}^2 = \alpha\mu_{ijk}$ is not likely to be valid under many circumstances and so the validity of the chi square values corrected by $1/\alpha$ is not determined by these considerations. It nevertheless appears to be advisable to calculate these corrected chi square values in order to gain experience concerning their usefulness in model selection.

APPENDIX B

FEASIBLE AND INFEASIBLE MARGINAL CONDITIONS

As noted in Section 3.3.2 a requirement for the existence of a solution to any of the three equivalent formulations of IPF (and hence for the existence of a unique solution to all three) is the existence of a feasible solution to the marginal and positivity conditions:

$$X_{ij+} = R_{ij}$$

$$X_{++k} = S_k$$

and

$$X_{ijk} > 0$$

in the case of the usual example.

The margins are said to be compatible if they agree along all common lower order margins e.g. R_{ij} and S_k are compatible if $R_{++} = S_+$ (R_{ij} and T_{jk} would be compatible if $R_{+j} = T_{j+}$). It is a perhaps surprising fact that there are compatible sets of (multi-dimensional) margins which are nevertheless infeasible. As noted in Section 3.3.5 this leads to a particularly intractable situation which cannot be dealt with satisfactorily using IPF. The margins in this case are essentially inconsistent and the situation can be referred to as essential infeasibility. As noted in Section 3.3.5, the practical technique for detecting essential infeasibility recommended in this report is to attempt to apply IPF and check for stable cycling with the developent of zero cells in the matrix being computed (see Section 3.3.5 for details). However, it is of some interest to note that at least in special cases explicit conditions for essential infeasibility can be given.

Note that the marginal conditions always have some solution (assuming they are compatible) and infeasibility arises only when any solution must have negative elements (of the X_{ijk} matrix).

Darroch (reference 5)considered the 2X2X2 case (three variables each with two levels) and found that the following conditions are necessary and sufficient for the existence of a positive solution. Suppose P_{++1}, P_{1++}, P_{+11}, etc. stand for specific positive margins of a positive matrix P_{ijk} ($_{ijk}$ each have 2 levels) then:

$P_{+++} - P_{1++} - P_{+1+} - P_{++1} + P_{+11} + P_{1+1} + P_{11+} > 0$

$P_{+11} - P_{1+1} - P_{11+} + P_{1++} > 0$

$P_{1+1} - P_{11+} - P_{+11} + P_{+1+} > 0$

$P_{11+} - P_{+11} - P_{1+1} + P_{++1} > 0$

$P_{+jk} > 0,\ P_{i+k} > 0,\ P_{ij+} > 0$

are the conditions.

A set of three compatible all-positive two dimensional margins which do not satisfy these conditions is easy to find. The following are an example:

$$X_{ij+} = \begin{pmatrix} 1 & 3 \\ 2 & 1 \end{pmatrix}$$

$$X_{i+k} = \begin{pmatrix} 3 & .5 \\ 1 & 2.5 \end{pmatrix}$$

$$X_{+jk} = \begin{pmatrix} 2 & 1.5 \\ 1 & 2.5 \end{pmatrix}$$

When a core matrix of all ones (it could be any positive core to illustrate the point) is scaled to these margins, convergence does not take place and instead stable cycling is observed. In addition, as expected from the discussion in Section 3.3.5 zeros develop in the matrices produced while attempting IPF. The matrices

$$X_{ijk}^{(3n+1)},\ X_{ijk}^{(3n+2)},\ X_{ijk}^{(3n+3)}$$

representing the three stages in a cycle are:

$$\begin{pmatrix} 1.000 & 1.01199 \\ 1.04853 & 0.000 \\ 0.000 & 1.99801 \\ 0.95147 & 1.000 \end{pmatrix}$$

$$\begin{pmatrix} 1.46447 & 0.500 \\ 1.53553 & 0.000 \\ 0.000 & 1.66332 \\ 1.000 & 0.83668 \end{pmatrix}$$

$$\begin{pmatrix} 1.49046 & 0.50904 \\ 1.500 & 0.000 \\ 0.000 & 1.000 \\ 1.36115 & 1.13885 \end{pmatrix}$$

APPENDIX C
LISTING OF PROGRAMS

The following program listingas are included in this report:

The first and last programs are well checked out but the program EM.FOR is not.

```
C     PROGRAM PROCAL.FOR      EDITED JUNE 22, 1982
C     PROGRAM TO DO ITERATIVE PROPORTIONAL SCALING.
C     THE PROGRAM READS FROM A FILE A CUBE MATRIX WITH THE MAXIMUM NUMBER OF
C     DIMENSIONS 5 AND A MAXIMUM NUMBER OF 5 LEVELS PER DIMENSION.
C     MARGINAL MATRICES OF LESSER DIMENSIONS THAN THE CUBE ARE
C     READ FROM A SECOND FILE, AND THE CUBE IS ITERATIVELY SCALED TO
C     THE MARGINALS.  THE PROGRAM DOES NOT CHECK THAT THE MARGINALS
C     ARE CONSISTENT.
C
C                DEFINITION OF VARIABLES
C                      (TO BE FILLED IN)
C
      INTEGER CORNUM(10,5),ENDFLG
      DIMENSION CUBE(5,5,5,5,5),LEVCUR(0/5),MARDIM(10)
      DIMENSION CUMULD(5,5,5,5,5)
      DIMENSION N(5),M(4),CUMARG(5,5,5,5)
      REAL MARG(10,5,5,5,5)
      DOUBLE PRECISION CUBFIL,MARFIL
      DATA LEVCUR(0)/1/
C
      IREAD  = 5
      IREAD1 = 20
      IREAD2 = 21
C
      ITYPE  = 5
      IWRITE = 5
C
      WRITE(ITYPE,22222)
22222 FORMAT(1X,'TYPE IN 1 IF YOU WISH OUTPUT TO BE PRINTED '/
     1 'ON FILE OUTPUT.DAT',//,1X,
     2 'TYPE IN 0 IF YOU WISH OUTPUT TYPED ON TTY')
      READ(5,22) IWRITE
C
      IF(IWRITE .EQ. 0) GO TO 5
      IWRITE = 22
C
      OPEN(UNIT=22,DEVICE='DSK',FILE='OUTPUT.DAT',ACCESS='SEQOUT')
C
    5 CONTINUE
C
      WRITE(ITYPE,888)
  888 FORMAT(1X,'TYPE IN 1 IF YOU WISH INPUT DATA TO BE PRINTED',//
     1 ,1X,'OTHERWISE TYPE IN 0')
      READ(IREAD,22) IPRINT
      WRITE(ITYPE,999)
  999 FORMAT(1X,'TYPE IN 1 IF YOU WISH CUMARG TO BE PRINTED',//
     1 ,1X,'OTHERWISE TYPE IN 0')
      READ(IREAD,22) KPRINT
C
C     INITIALIZATION OF LEVCUR
C
      DO 10 I = 1,5
            LEVCUR(I) = 1
   10 CONTINUE
C
C     INITIALIZATION OF CORNUM
C
      DO 20 I = 1,10
            DO 30 J = 1,5
                  CORNUM(I,J) = 0
   30       CONTINUE
```

```
   20   CONTINUE
C
C                    OBTAIN USER INPUT
C
C
C       READ IN NAME OF CORE MATRIX FROM THE TTY
C
        WRITE(ITYPE,111)
  111   FORMAT(5X,'NAME OF CORE MATRIX FILE:',5X,$)
        READ(IREAD,11) CORFIL
   11   FORMAT(A10)
C
        OPEN(UNIT=IREAD1,DEVICE='DSK',FILE=CORFIL,ACCESS='SEQIN')
C
C       GET INFORMATION ON CORE MATRIX FROM FILE CORFIL
C
        READ(IREAD1,22) NDIM
   22   FORMAT(I012)
C
        READ(IREAD1,22) (LEVCOR(I),I=1,NDIM)
C
        IF(IPRINT .EQ. 0) GO TO 50
C
        WRITE(IRITE,1111) NDIM
 1111   FORMAT(1X,'NDIM=',I3,///)
        WRITE(IRITE,2222) (I,LEVCOR(I),I=1,NDIM)
 2222   FORMAT(1X,'NUMBER OF LEVELS FOR DIMENSION ',I2,' = ',I2)
C
   50   CONTINUE
C
C       READ IN CORE MATRIX FROM FILE CORFIL AND WRITE OUT ON TTY
C
        IF(IPRINT .EQ. 0) GO TO 60
C
        WRITE(IRITE,3333)
 3333   FORMAT(///,20X,'CORE MATRIX',//)
C
   60   CONTINUE
C
        DO 100 N5=1,LEVCOR(5)
               DO 110 N4=1,LEVCOR(4)
                      DO 120 N3=1,LEVCOR(3)
                             DO 130 N2=1,LEVCOR(2)
                                    READ(IREAD1,33)(CORE(N1,N2,N3,N4,N5)
     1                              ,N1=1,LEVCOR(1))
   33                               FORMAT(5F)
C
                                    IF(IPRINT.EQ.0)GO TO 130
C
                                    WRITE(IRITE,4444) (N1,N2,N3,N4,N5,
     1                              CORE(N1,N2,N3,N4,N5),
     2                              N1=1,LEVCOR(1))
 4444                               FORMAT(1X,5(3I2,F13.5,2X))
  130                          CONTINUE
  120                   CONTINUE
  110            CONTINUE
  100   CONTINUE
C
C       READ IN NAME OF MARGINAL MATRIX FROM THE TTY
C
        WRITE(ITYPE,222)
  222   FORMAT(///,5X,'NAME OF MARGINAL MATRICES FILE:',5X,$)
```

```
      READ(IREAD,11) MARFIL
C
      OPEN(UNIT=IREAD2,DEVICE='DSK',FILE=MARFIL,ACCESS='SEQIN')
C
C     GET INFORMATION ON MARGINAL MATRICES FROM FILE MARFIL
C
      READ(IREAD2,22) NMARG
C
      IF(IPRINT .EQ. 0) GO TO 150
C
      WRITE(IWRITE,5555) NMARG
 5555 FORMAT(1X,'NUMBER OF MARGINAL MATRICES=',I3)
C
  150 CONTINUE
C
      DO 200 I=1,NMARG
            READ(IREAD2,22) MARDIM(I)
C
            IF(IPRINT .EQ.0) GO TO 170
C
            WRITE(IWRITE,6666) I,MARDIM(I)
 6666       FORMAT(1X,'NUMBER OF DIMENSIONS OF MARGINAL ',I2,' =',I2)
C
  170 CONTINUE
C
            READ(IREAD2,22) (CURRDM(I,J),J=1,MARDIM(I))
C
      IF(IPRINT .EQ. 0) GO TO 200
C
            WRITE(IWRITE,7777) (I,J,CURRDM(I,J),J=1,MARDIM(I))
 7777       FORMAT(1X,'CURRDM',1X,5(2I3,3X,I2,6X))
C
  200 CONTINUE
C
C     READ IN MARGINALS FROM FILE MARFIL AND WRITE OUT ON TTY
C
      IF(IPRINT .EQ. 0) GO TO 250
C
      WRITE(IWRITE,8888)
 8888 FORMAT(///,20X,'MARGINALS',///)
C
  250 CONTINUE
C
      DO 300 J=1,NMARG
            DO 310 M4=1,LEVCUR(CURRDM(J,4))
                  DO 320 M3=1,LEVCUR(CURRDM(J,3))
                        DO 330 M2=1,LEVCUR(CURRDM(J,2))
                              READ(IREAD2,33)(MARG(J,M1,M2,M3,M4),
     1                        M1=1,LEVCUR(CURRDM(J,1)))
C
                              IF(IPRINT.EQ.0)GO TO 330
C
                              WRITE(IWRITE,4444) (J,M1,M2,M3,M4,
     1                        MARG(J,M1,M2,M3,M4),
     2                        M1=1,LEVCUR(CURRDM(J,1)))
  330                   CONTINUE
  320             CONTINUE
  310       CONTINUE
  300 CONTINUE
C
C   COMPUTE MARGINALS FROM CORE AND SCALE CORE
```

```
C   CONTINUE UNTIL CONVERGENCE OR MAXIMUM NUMBER OF ITERATIONS
C   HAS BEEN REACHED.  EACH ITERATION LOOPS THRU ALL MARGINALS
C
      WRITE(ITYPE,600)
  600 FORMAT(///,1X,'ENTER MAXIMUM NUMBER OF ITERATIONS',5X,$)
      READ(IREAD,60) MAXIT
   60 FORMAT(I5)
      WRITE(ITYPE,777)
  777 FORMAT(///,1X,'ENTER EPSIL',5X,$)
      READ(IREAD,77) EPSIL
   77 FORMAT(F10.3)
C
      WRITE(ITYPE,55555) MAXIT,EPSIL
55555 FORMAT(1H1,////,' MAXIT=',I5,5X,'EPSIL=',E15.7,///)
C
      ITER=0
  400 ENDFLG=1
      ITER=ITER+1
C
      IF(ITER .EQ. 1) ENDFLG = 0
C
      DIFMAX = 0.
      N1DIF  = 0
      N2DIF  = 0
      N3DIF  = 0
      N4DIF  = 0
      N5DIF  = 0
C
      RATMAX = 0.
      N1RAT  = 0
      N2RAT  = 0
      N3RAT  = 0
      N4RAT  = 0
      N5RAT  = 0
C
      DELMAX = 0.
      N1DEL  = 0
      N2DEL  = 0
      N3DEL  = 0
      N4DEL  = 0
      N5DEL  = 0
C
      PRUMAX = 0.
      N1PRU  = 0
      N2PRU  = 0
      N3PRU  = 0
      N4PRU  = 0
      N5PRU  = 0
C
      DO 600 J=1,NMARG
C
C   REINITIALIZE COMPUTED MARGINAL MATRIX
C
      DO 410 I1=1,5
            DO 420 I2=1,5
                  DO 430 I3=1,5
                        DO 440 I4=1,5
                              COMARG(I1,I2,I3,I4)=0.
  440                   CONTINUE
  430             CONTINUE
  420       CONTINUE
```

```
  410   CONTINUE
C
      IF(MARDIM(J).EQ.4)GOTO 460
C
C     SET UNUSED DIMENSION INDICES OF MARGINALS TO 1
C
      DO 450 K=MARDIM(J)+1,4
           M(K)=1
  450 CONTINUE
C
C   SET POSSIBLY UNUSED DIMENSION OF CUBE INDICES TO 1
C
  460 DO 470 K=J,5
           N(K)=1
  470 CONTINUE
C
C   LOOP THRU ALL CUBE VALUES AND COMPUTE MARGINAL VALUES FROM CUBE
C
      IF(KPRINT .EQ. 0) GO TO 475
C
      WRITE(IWRITE,11111)
11111 FORMAT(///,20A,'CUMARG',///)
C
  475 CONTINUE
C
      DO 480 N5=1,LEVCUR(5)
          N(5)=N5
          DO 490 N4=1,LEVCUR(4)
              N(4)=N4
              DO 500 N3=1,LEVCUR(3)
                  N(3)=N3
                  DO 510 N2=1,LEVCUR(2)
                      N(2)=N2
                      DO 520 N1=1,LEVCUR(1)
                          N(1)=N1
                          DO 530 K=1,MARDIM(J)
                              M(K)=N(CURKUM(J,K))
  530                     CONTINUE
                          CUMARG(M(1),M(2),M(3),M(4))
     1                    =CUMARG(M(1),M(2),M(3),M(4))+
     2                    CUBE(N1,N2,N3,N4,N5)
                          IF(KPRINT.EQ.0) GO TO 520
                          WRITE(IWRITE,4444)
     1                    M(1),M(2),M(3),M(4),
     2                    CUMARG(M(1),M(2),M(3),M(4))
  520                 CONTINUE
  510             CONTINUE
  500         CONTINUE
  490     CONTINUE
  480 CONTINUE
C
C   SCALE CUBE BY GIVEN MARGINAL DIVIDED BY COMPUTED MARGINAL
C
      DO 540 N5=1,LEVCUR(5)
          N(5)=N5
          DO 550 N4=1,LEVCUR(4)
              N(4)=N4
              DO 560 N3=1,LEVCUR(3)
                  N(3)=N3
                  DO 570 N2=1,LEVCUR(2)
                      N(2)=N2
```

```
      DO 580 N1=1,LEVCUR(1)
         M(1)=N1
         DO 590 K=1,MARGIN(J)
            M(K)=N(CURRUN(J,K))
  590    CONTINUE
         SFACT=MARG(J,M(1),M(2),
     1   M(3),M(4))/CUMARG(
     2   M(1),M(2),M(3),M(4))
         CURVAL=CURF(N1,N2,N3,N4,N5)
         CURNEW=CURVAL*SFACT
         DIFF=ABS(CURNEW-CURVAL)
         IF(DIFF.LE.DIFMAX)GO TO 592
         DIFMAX=DIFF
         N1DIF=N1
         N2DIF=N2
         N3DIF=N3
         N4DIF=N4
         N5DIF=N5
C
  592    CONTINUE
         RATIO=CURNEW/CURVAL
         IF(RATIO.LE.RATMAX)GO TO 594
         RATMAX=RATIO
         N1RAT=N1
         N2RAT=N2
         N3RAT=N3
         N4RAT=N4
         N5RAT=N5
C
  594    CONTINUE
         IF(J.NE.NMARG)GO TO 599
C
         IF(ITER.GT.1)GO TO 595
         CUROLD(N1,N2,N3,N4,N5)=CURNEW
         GO TO 599
C
  595    CONTINUE
         DIFF=ABS(CURNEW-CUROLD(N1,N2,N3,N4,N5))
         IF(DIFF.LE.DELMAX)GO TO 596
         DELMAX=DIFF
         N1DEL=N1
         N2DEL=N2
         N3DEL=N3
         N4DEL=N4
         N5DEL=N5
C
  596    CONTINUE
         RATIO=CURNEW/CUROLD(N1,N2,N3,N4,N5)
         IF(RATIO.LE.PRUMAX)GO TO 598
         PRUMAX=RATIO
         N1PRU=N1
         N2PRU=N2
         N3PRU=N3
         N4PRU=N4
         N5PRU=N5
C
  598    CONTINUE
         IF(ABS(CURNEW-CUROLD(N1,N2,N3,N4,N5))
     1   .LE.EPSIL)GO TO 597
         ENDFLG=0
C
```

```
  547                        CONTINUE
                             CUROLD(N1,N2,N3,N4,N5)=CURNEW
C
  599                        CONTINUE
                             CURE(N1,N2,N3,N4,N5)=CURNEW
  580                      CONTINUE
  570                    CONTINUE
  560                  CONTINUE
  550                CONTINUE
  540          CONTINUE
  600          CONTINUE
               IF(IERDFLG.EQ.1)GOTO 700          !HAVE FINAL CURE MATRIX
               IF(ITER.LT.MAXIT)GO TO 400        !MAXIMUM # OF ITERATIONS NOT REACHED
C
C    MAXIMUM NUMBER OF ITERATIONS REACHED OR METHOD CONVERGES
C
  700          CONTINUE
C
C    TYPE OUT FINAL CURE VALUES
C
               WRITE(IWRITE,4411)
 4411          FORMAT(20X,'COMPARISON OF FINAL MARGINALS',
     1          ' FOR TWO CONSECUTIVE ITERATIONS')
               WRITE(IWRITE,4422) ITER
 4422          FORMAT(/,5X,'NUMBER OF ITERATIONS=',I5)
               WRITE(IWRITE,4433) DELMAX,N1DEL,N2DEL,N3DEL,N4DEL,N5DEL
 4433          FORMAT(/,5X,'ABSOLUTE VALUE OF MAX DEVIATION IS ',E15.7,
     1          ' AT POINT ',5I3)
               WRITE(IWRITE,4455) PROMAX,N1PRO,N2PRO,N3PRO,N4PRO,N5PRO
 4455          FORMAT(/,5X,'MAX RATIO IS ',E15.7,' AT POINT ',5I3)
C*******
               WRITE(3,4411)
               WRITE(3,4422) ITER
               WRITE(3,4433) DELMAX,N1DEL,N2DEL,N3DEL,N4DEL,N5DEL
               WRITE(3,4455) PROMAX,N1PRO,N2PRO,N3PRO,N4PRO,N5PRO
C*******
               WRITE(IWRITE,4466)
 4466          FORMAT(////,20X,'COMPARISON OF TWO CONSECUTIVE MARGINALS')
               WRITE(IWRITE,4433) DIFMAX,N1DIF,N2DIF,N3DIF,N4DIF,N5DIF
               WRITE(IWRITE,4455) RATMAX,N1RAT,N2RAT,N3RAT,N4RAT,N5RAT
C*******
               WRITE(3,4466)
               WRITE(3,4433) DIFMAX,N1DIF,N2DIF,N3DIF,N4DIF,N5DIF
               WRITE(3,4455) RATMAX,N1RAT,N2RAT,N3RAT,N4RAT,N5RAT
C*******
C
               WRITE(IWRITE,4477)
 4477          FORMAT(1H1,////)
C
               DO 710 N5=1,LEVCUR(5)
                       DO 720 N4=1,LEVCUR(4)
                               DO 730 N3=1,LEVCUR(3)
                                       DO 740 N2=1,LEVCUR(2)
                                               WRITE(IWRITE,4444) (N1,N2,N3,N4,N5,
     1                                         CURE(N1,N2,N3,N4,N5),
     2                                         N1=1,LEVCUR(1))
  740                                  CONTINUE
  730                          CONTINUE
  720                  CONTINUE
  710          CONTINUE
               CALL EXIT
```

```
      PROGRAM EM
C     PROGRAM EM.FOR              EDITED OCTOBER 7, 1982
C     PROGRAM TO DO ITERATIVE WEIGHTED SCALING.
C     THE PROGRAM READS FROM A FILE A CORE MATRIX WITH THE MAXIMUM NUMBER OF
C     5 DIMENSIONS AND A MAXIMUM NUMBER OF 5 LEVELS PER DIMENSION.
C     MARGINAL MATRICES OF LESSER DIMENSIONS THAN THE CORE ARE
C     READ FROM A SECOND FILE, AND THE CORE IS ITERATIVELY SCALED TO
C     A WEIGHTED COMBINATION OF THE MARGINALS.
C     THE PROGRAM DOES NOT CHECK THAT THE MARGINALS ARE CONSISTENT.
C     ARE CONSISTENT.
C
      INTEGER CORROW(10,5),ENDFLG
      DIMENSION CORE(5,5,5,5,5),LEVCOR(0/5),MARDIM(10)
      DIMENSION N(5),M(4),COMARG(5,5,5,5),WEIGHT(0/10)
      REAL MARG(10,5,5,5,5)
      DIMENSION TMPCOR(5,5,5,5,5),CORNEW(5,5,5,5,5)
      DOUBLE PRECISION COREFIL,MAREFIL,OUTFIL
      DATA LEVCOR(0)/1/
C
      IREAD  = 5
      IREAD1 = 20
      IREAD2 = 21
C
      ITYPE  = 5
      IPITE  = 5
C
      WRITE(ITYPE,22222)
22222 FORMAT(1X,'TYPE IN 1 IF YOU WISH OUTPUT TO BE PRINTED '
     1 'ON A FILE',//,1X,
     2 'TYPE IN 0 IF YOU WISH OUTPUT TYPED ON TTY')
      READ(5,22) IWRITE
   22 FORMAT(I0I2)
C
      IF(IWRITE .EQ. 0) GO TO 5
      IRITE = 22
      WRITE(ITYPE,1144)
 1144 FORMAT(2X,'TYPE NAME OF OUTPUT FILE',5X,$)
      READ(IREAD,1155) OUTFIL
 1155 FORMAT(A10)
C
      OPEN(UNIT=22,DEVICE='DSK',FILE=OUTFIL,ACCESS='SEQOUT')
C
    5 CONTINUE
C
      WRITE(ITYPE,888)
  888 FORMAT(1X,'TYPE IN 1 IF YOU WISH INPUT DATA TO BE PRINTED',/
     1 ,1X,'OTHERWISE TYPE IN 0')
      READ(IREAD,22) IPRINT
      WRITE(ITYPE,999)
  999 FORMAT(1X,'TYPE IN 1 IF YOU WISH COMARG TO BE PRINTED',/
     1 ,1X,'OTHERWISE TYPE IN 0')
      READ(IREAD,22) KPRINT
C
C     INITIALIZATION OF LEVCOR
C
      DO 10 I = 1,5
            LEVCOR(I) = 1
   10 CONTINUE
C
C     INITIALIZATION OF CORRO
C
```

```
      DO 20 I = 1,10
            DO 30 J = 1,5
                  CORROW(I,J) = 0
   30       CONTINUE
   20 CONTINUE
C
C                 OBTAIN USER INPUT
C
C     READ IN NAME OF CORE MATRIX FROM THE TTY
C
      WRITE(ITYPE,111)
  111 FORMAT(5X,'NAME OF CORE MATRIX FILE:',5X,$)
      READ(IREAD,11) CORFIL
   11 FORMAT(A10)
C
      OPEN(UNIT=IREAD1,DEVICE='DSK',FILE=CORFIL,ACCESS='SEQIN')
C
C     GET INFORMATION ON CORE MATRIX FROM FILE CORFIL
C
      READ(IREAD1,1122) NDIM,WEIGHT(0)
 1122 FORMAT(I2,F10.3)
C
      READ(IREAD1,22) (LEVCOR(I),I=1,NDIM)
C
      IF(IPRINT .EQ. 0) GO TO 50
C
      WRITE(IRITE,1111) NDIM,WEIGHT(0)
 1111 FORMAT(1X,'NDIM=',I3,3X,'WEIGHT(0)=',E15.7,//)
      WRITE(IRITE,2222) (I,LEVCOR(I),I=1,NDIM)
 2222 FORMAT(1X,'NUMBER OF LEVELS FOR DIMENSION ',I2,' = ',I2)
C
   50 CONTINUE
C
C     READ IN CORE MATRIX FROM FILE CORFIL AND WRITE OUT ON TTY
C
      IF(IPRINT .EQ. 0) GO TO 60
C
      WRITE(IRITE,3333)
 3333 FORMAT(1H1,//,16X,'INITIAL CORE MATRIX',//)
C
   60 CONTINUE
C
C     READ IN CORE MATRIX AND INITIALIZE CORNEW
C
      DO 100 N5=1,LEVCOR(5)
          DO 110 N4=1,LEVCOR(4)
              DO 120 N3=1,LEVCOR(3)
C                 READ IN CORE MATRIX
                  DO 130 N2=1,LEVCOR(2)
                      READ(IREAD1,33)(CORE(N1,N2,N3,N4,N5)
     1                ,N1=1,LEVCOR(1))
   33                 FORMAT(5E)
C                     INITIALIZE CORNEW
                      DO 140 N1=1,LEVCOR(1)
                          CORNEW(N1,N2,N3,N4,N5)=CORE(N1,N2,N3,N4,N5)
  140                 CONTINUE
C
                      IF(IPRINT.EQ.0)GO TO 130
C
                      WRITE(IRITE,4444) (N1,N2,N3,N4,N5,
     1                CORE(N1,N2,N3,N4,N5),
```

```
                M1=1,LEVCOR(1))
 4444           FORMAT(1X,5(5I2,F13.5,2X))
  130         CONTINUE
  120       CONTINUE
  110     CONTINUE
  100   CONTINUE
C
C       READ IN NAME OF MARGINAL MATRIX FROM THE TTY
C
        WRITE(ITYPE,222)
  222   FORMAT(//,5X,'NAME OF MARGINAL MATRICES FILE:',5X,$)
        READ(IREAD,11) MARFIL
C
        OPEN(UNIT=IREAD2,DEVICE='DSK',FILE=MARFIL,ACCESS='SEQIN')
C
C       GET INFORMATION ON MARGINAL MATRICES FROM FILE MARFIL
C
        READ(IREAD2,22) NMARG
C
        IF(IPRINT .EQ. 0) GO TO 150
C
        WRITE(IRITE,5555) NMARG
 5555   FORMAT(1H1,//,1X,'NUMBER OF MARGINAL MATRICES=',I3,//)
C
  150   CONTINUE
C
        DO 200 I=1,NMARG
                READ(IREAD2,1122) MARDIM(I),WEIGHT(I)
C
                IF(IPRINT .EQ.0) GO TO 170
C
                WRITE(IRITE,6666) I,MARDIM(I),I,WEIGHT(I)
 6666           FORMAT(1X,'NUMBER OF DIMENSIONS OF MARGINAL ',I2,' =',I2
     1          ,1X,'WEIGHT(',I2,')=',F15.7)
C
  170   CONTINUE
C
                READ(IREAD2,22) (CORRDM(I,J),J=1,MARDIM(I))
C
        IF(IPRINT .EQ. 0) GO TO 200
C
                WRITE(IRITE,7777) (I,J,CORRDM(I,J),J=1,MARDIM(I))
 7777           FORMAT(1X,'CORRDM',1X,5(2I3,3X,I2,6X))
C
  200   CONTINUE
C
C       READ IN MARGINALS FROM FILE MARFIL AND WRITE OUT ON TTY
C
        IF(IPRINT .EQ. 0) GO TO 250
C
        WRITE(IRITE,8888)
 8888   FORMAT(///,20X,'MARGINALS',//)
C
  250   CONTINUE
C
        DO 300 J=1,NMARG
                DO 310 M4=1,LEVCOR(CORRDM(J,4))
                    DO 320 M3=1,LEVCOR(CORRDM(J,3))
                            DO 330 M2=1,LEVCOR(CORRDM(J,2))
                                    READ(IREAD2,33)(MARG(J,M1,M2,M3,M4),
     1                              M1=1,LEVCOR(CORRDM(J,1)))
```

```
C
                                        IF(IPRINT.EQ.0)GO TO 330
C
                                        WRITE(IPITE,4444) (J,M1,M2,M3,M4,
     1                                  MARG(J,M1,M2,M3,M4),
     2                                  M1=1,LEVCOR(CORRDM(J,1)))
  330                           CONTINUE
  320                     CONTINUE
  310               CONTINUE
  300     CONTINUE
C
C    COMPUTE MARGINALS FROM CORE AND SCALE CORE
C    CONTINUE UNTIL CONVERGENCE OR MAXIMUM NUMBER OF INTERATIONS
C    HAS BEEN REACHED.  EACH ITERATION LOOPS THRU ALL MARGINALS
C
          WRITE(ITYPE,666)
  666     FORMAT(//,1X,'ENTER MAXIMUM NUMBER OF ITERATIONS',5X,$)
          READ(IREAD,66) MAXIT
   66     FORMAT(I5)
          WRITE(ITYPE,777)
  777     FORMAT(//,1X,'ENTER EPSIL',5X,$)
          READ(IREAD,77) EPSIL
   77     FORMAT(F10.3)
C
          WRITE(IPITE,55555) MAXIT,EPSIL
55555     FORMAT(1H1,//,' MAXIT=',I5,5X,'EPSIL=',E15.7,//)
C
          ITER=0
  400     ENDFLG=1
          ITER=ITER+1
C
C
C    REINITIALIZE TEMPORARY CORE MATRIX
C
          DO 410 I1=1,5
              DO 412 I2=1,5
                  DO 414 I3=1,5
                      DO 416 I4=1,5
                          DO 418 I5=1,5
                          TMPCOR(I1,I2,I3,I4,I5)=0.
  418                     CONTINUE
  416                 CONTINUE
  414             CONTINUE
  412         CONTINUE
  410     CONTINUE
C
          DO 500 J=1,NMARG
C
C    REINITIALIZE COMPUTED MARGINAL MATRIX
C
          DO 420 I1=1,5
              DO 430 I2=1,5
                  DO 440 I3=1,5
                      DO 445 I4=1,5
                          CONMRG(I1,I2,I3,I4)=0.
  445                     CONTINUE
  440             CONTINUE
  430         CONTINUE
  420     CONTINUE
C
          IF(MARDIM(J).EQ.4)GOTO 460
```

```
C
C       SET UNUSED DIMENSION INDICES OF MARGINALS TO 1
C
        DO 460 K=MARDIM(J)+1,4
             N(K)=1
  460   CONTINUE
C
C   SET POSSIBLY UNUSED DIMENSION OF CORE INDICES TO 1
C
  460   DO 470 K=3,5
             N(K)=1
  470   CONTINUE
C
C   LOOP THRU ALL CORE VALUES AND COMPUTE MARGINAL VALUES FROM CORE
C
        IF(KPRINT .EQ. 0) GO TO 475
C
        WRITE(IRITE,11111)
11111   FORMAT(//,20X,'COMARG',///)
C
  475   CONTINUE
C
        DO 480 N5=1,LEVCOR(5)
            N(5)=N5
            DO 490 N4=1,LEVCOR(4)
                N(4)=N4
                DO 500 N3=1,LEVCOR(3)
                   N(3)=N3
                      DO 510 N2=1,LEVCOR(2)
                         N(2)=N2
                            DO 520 N1=1,LEVCOR(1)
                               N(1)=N1
                               DO 530 K=1,MARDIM(J)
                                  M(K)=N(CORRDN(J,K))
  530                          CONTINUE
                            COMARG(M(1),M(2),M(3),M(4))
     1                      =COMARG(M(1),M(2),M(3),M(4))+
     2                      CORNEW(N1,N2,N3,N4,N5)
                            IF(KPRINT.EQ.0) GO TO 520
                            WRITE(IRITE,44444)
     1                      M(1),M(2),M(3),M(4),
     2                      COMARG(M(1),M(2),M(3),M(4))
44444                       FORMAT(1X,5(4I2,E15.7,2X))
  520                    CONTINUE
  510                CONTINUE
  500            CONTINUE
  490        CONTINUE
  480    CONTINUE
C
C   SCALE CORE BY GIVEN MARGINAL DIVIDED BY COMPUTED MARGINAL
C
        DO 540 N5=1,LEVCOR(5)
            N(5)=N5
            DO 550 N4=1,LEVCOR(4)
                N(4)=N4
                DO 560 N3=1,LEVCOR(3)
                    N(3)=N3
                    DO 570 N2=1,LEVCOR(2)
                        N(2)=N2
                        DO 580 N1=1,LEVCOR(1)
                            N(1)=N1
```

```
                        DO 590 K=1,MAPDIM(J)
                            M(K)=M(CURMDM(J,K))
  590                   CONTINUE
                        SFACT=MARG(J,M(1),M(2), M(3),M(4))/
     1                  COMPRG(M(1),M(2),M(3),M(4))
                        TMPCOR(N1,N2,N3,N4,N5)=
     1                  TMPCOR(N1,N2,N3,N4,N5) +
     2                  WEIGHT(J)*SFACT*CORNEW(N1,N2,N3,N4,N5)
  580                 CONTINUE
  570               CONTINUE
  560             CONTINUE
  550           CONTINUE
  540         CONTINUE
C
  600         CONTINUE
C
              DIFMAX = 0.
C
              DO 640 N5=1,LEVCOR(5)
                  N(5)=N5
                  DO 650 N4=1,LEVCOR(4)
                      N(4)=N4
                      DO 660 N3=1,LEVCOR(3)
                          N(3)=N3
                          DO 670 N2=1,LEVCOR(2)
                              N(2)=N2
                              DO 680 N1=1,LEVCOR(1)
                                  N(1)=N1
                                  CORVAL=WEIGHT(0)*CORE(N1,N2,N3,N4,N5)+
     1                            TMPCOR(N1,N2,N3,N4,N5)
                                  DIFF=ABS(CORNEW(N1,N2,N3,N4,N5)-CORVAL)
                                  IF(DIFF.LE.DIFMAX) GO TO 675
                                  DIFMAX = DIFF
                                  INDX1  = N1
                                  INDX2  = N2
                                  INDX3  = N3
                                  INDX4  = N4
                                  INDX5  = N5
C
  675                             CONTINUE
                                  IF(DIFF.LE.EPSIL) GO TO 678
                                  ENDFLG = 0
C
  678                             CONTINUE
                                  CORNEW(N1,N2,N3,N4,N5)=CORVAL
  680                           CONTINUE
  670                       CONTINUE
  660                   CONTINUE
  650               CONTINUE
  640           CONTINUE
C
              IF(ENDFLG.EQ.1)GOTO 700          !HAVE FINAL CORE MATRIX
              IF(ITER.LT.MAXIT)GO TO 400       !MAXIMUM # OF ITERATIONS NOT REACHED
C
C   MAXIMUM NUMBER OF ITERATIONS REACHED OR METHOD CONVERGES
C
  700         CONTINUE
C
C   TYPE OUT FINAL CORE VALUES
C
              WRITE(IWRITE,4422) ITER,DIFMAX,INDX1,INDX2,INDX3,INDX4,INDX5
```

```
 4422   FORMAT(5X,'NUMBER OF ITERATIONS=',I5,
     1  5X,'MAX DEVIATION ',E15.7,2X,'OCCURS AT POINT (',
     2  4(I2,','),I2,')',//,16X,'FINAL CORE MATRIX',//)
C
        DO 710 N5=1,LEVCOR(5)
               DO 720 N4=1,LEVCOR(4)
                      DO 730 N3=1,LEVCOR(3)
                             DO 740 N2=1,LEVCOR(2)
                                    WRITE(IRITE,4444) (N1,N2,N3,N4,N5,
     1                              CORNEW(N1,N2,N3,N4,N5),
     2                              N1=1,LEVCOR(1))
  740                        CONTINUE
  730                 CONTINUE
  720           CONTINUE
  710   CONTINUE
        CALL EXIT
        END
```

```
      PROGRAM LOGLIN
C     PROGRAM LOGLIN          STUPED ON FILE LOGLIN.FOR
C     CREATED AUGUST 16, 1982          REVISED AUGUST 25, 1982
C
C     PROGRAM TO INPUT CORE MATRIX AND CREATE DESIRED MARGINALS (MARGEN)
C     USING THESE MARGINALS TO OBTAIN NEW CORE MATRIX (PRSCAL)
C     OBTAIN STATISTICS COMPARING OLD AND NEW CORE MATRICES (CHISQR)
C
C     MARGINAL INFORMATION IS READ  FROM A FILE DESIGNATED BY THE USER
C     THE MAIN PROGRAM READS IN THE FIRST CORE MATRIX
C     THIS MATRIX HAS A MAXIMUM OF 5 DIMENSIONS AND A MAXIMUM NUMBER
C     OF 5 LEVELS PER EACH DIMENSION
C
      INTEGER CORFDM
C
      REAL MARG
C
      DOUBLE PRECISION CORFIL,CORINI,INMARG,OUTFIL
C
      COMMON /CORES1/ CORFST(5,5,5,5,5)
      COMMON /CORES2/ CORNEW(5,5,5,5,5)
      COMMON /MARGIN/ NDIM,LEVCOR(0:5),NMARG,MARDIM(10),CORFDM(10,5),
     1 MARG(10,5,5,5,5,5)
      COMMON/INOUT/ IREAD,IREAD1,IREAD2,IREAD3,ITYPE,IRITE,IRITE1,
     1 IRITE2
      COMMON /PRINT/ IPRINT,KPRINT
      COMMON/ISCALE/ SCALE
C
      DATA LEVCOR(0)/1/
C
      IREAD  = 5
      IREAD1 = 20
      IREAD2 = 21
      IREAD3 = 22
C
      ITYPE  = 5
C
      IRITE  = 5
      IRITE1 = 24
      IRITE2 = 25
C
      WRITE(ITYPE,111)
  111 FORMAT(1X,'TYPE IN 1 IF YOU WISH OUTPUT TO BE PRINTED ',
     1 'ON A FILE',//,1X,
     2 'TYPE IN 0 IF YOU WISH OUTPUT TYPED ON TTY')
      READ(IREAD,11) IWRITE
   11 FORMAT(I0I2)
C
      IF(IWRITE .EQ. 0) GO TO 5
      IRITE = 23
      WRITE(ITYPE,222)
  222 FORMAT(1X,'TYPE NAME OF OUTPUT FILE',5X,$)
      READ(IREAD,22) OUTFIL
   22 FORMAT(A10)
C
C     OPEN OUTPUT FILE OUTFIL
C
      OPEN(UNIT=IRITE,DEVICE='DSK',FILE=OUTFIL,ACCESS='SEQOUT')
C
    5 CONTINUE
C
```

```
      WRITE(ITYPE,333)
  333 FORMAT(1X,'TYPE IN 1 IF YOU WISH INPUT DATA TO BE PRINTED',/
     1 ,1X,'OTHERWISE TYPE IN 0')
      READ(IREAD,11) IPRINT
C
C     INITIALIZE MARGINAL MATRIX MARG
C
      DO 700 I5= 1,5
         DO 710 I4 = 1,5
            DO 720 I3 = 1,5
               DO 730 I2 = 1,5
                  DO 740 I1 = 1,10
                     MARG(I1,I2,I3,I4,I5) = 0.
  740             CONTINUE
  730          CONTINUE
  720       CONTINUE
  710    CONTINUE
  700 CONTINUE
C
C     INITIALIZATION OF LEVCOR
C
      DO 10 I = 1,5
         LEVCOR(I) = 1
   10 CONTINUE
C
C     INITIALIZATION OF CORRDM
C
      DO 20 I = 1,10
         DO 30 J = 1,5
            CORRDM(I,J) = 0
   30    CONTINUE
   20 CONTINUE
C
C                OBTAIN USER INPUT
C
C     READ IN SCALING FACTOR TO MAKE OUTPUT MORE READABLE
C
      WRITE(ITYPE,777)
  777 FORMAT(1X,'ENTER SCALING FACTOR IN F FORMAT')
      READ(IREAD,44) SCALE
   44 FORMAT(F)
C
C     READ IN NAME OF MARGINAL INFORMATION FILE
C
      WRITE(ITYPE,444)
  444 FORMAT(1X,'TYPE NAME OF MARGINAL INFORMATION FILE',5X,$)
      READ(IREAD,22) INMARG
C
C     READ IN NAME OF FIRST CORE MATRIX FILE FROM THE TTY
C
      WRITE(ITYPE,555)
  555 FORMAT(1X,'TYPE NAME OF FIRST CORE MATRIX FILE',5X,$)
      READ(IREAD,22) COREIL
C
C     READ IN NAME OF FILE THAT INITIALIZES THE CORE MATRIX
C           FOR SUBROUTINE PRSCAL
C
      WRITE(ITYPE,666)
  666 FORMAT(1X,'TYPE FILE NAME FOR INITIAL CORE MATRIX FOR PRSCAL',
     1 5X,$)
      READ(IREAD,22) COREINI
```

```
C
C       OPEN MARGINAL INFORMATION FILE INMARG
C
        OPEN(UNIT=IREAD1,DEVICE='SCIA',FILE=INMARG,ACCESS='SEQIN')
C
C       GET CORE INFORMATION FROM FILE INMARG
C
        READ(IREAD1,11) NDIM
C
        READ(IREAD1,11) (LEVCOR(I),I=1,NDIM)
C
        IF(IPRINT .EQ. 0) GO TO 50
C
        WRITE(IRITE,1111) NDIM
 1111   FORMAT(1X,'NDIM=',I3,//)
        WRITE(IRITE,2222) (I,LEVCOR(I),I=1,NDIM)
 2222   FORMAT(1X,'NUMBER OF LEVELS FOR DIMENSION ',I2,' = ',I2)
C
   50   CONTINUE
C
C       GET INFORMATION ON MARGINAL MATRICES FROM FILE INMARG
C
        READ(IREAD1,11) NMARG
C
        IF(IPRINT .EQ. 0) GO TO 60
C
        WRITE(IRITE,3333) NMARG
 3333   FORMAT(1X,'NUMBER OF MARGINAL MATRICES=',I3)
C
   60   CONTINUE
C
        DO 100 I=1,NMARG
                READ(IREAD1,11) MARDIM(I)
C
                IF(IPRINT .EQ.0) GO TO 130
C
                WRITE(IRITE,4444) I,MARDIM(I)
 4444           FORMAT(1X,'NUMBER OF DIMENSIONS OF MARGINAL ',I2,' =',I2)
C
  130   CONTINUE
C
                READ(IREAD1,11) (CORRDM(I,J),J=1,MARDIM(I))
C
        IF(IPRINT .EQ. 0) GO TO 100
C
                WRITE(IRITE,5555) (I,J,CORRDM(I,J),J=1,MARDIM(I))
 5555           FORMAT(1X,'CORRDM',1X,5(2I3,3X,I2,6X))
C
  100   CONTINUE
C
C       CLOSE FILE INMARG
C
        CLOSE(UNIT=IREAD1)
C
C       OPEN FIRST CORE MATRIX FILE CORFIL
C
        OPEN(UNIT=IREAD2,DEVICE='SCIA',FILE=CORFIL,ACCESS='SEQIN')
C
C       READ IN FIRST CORE MATRIX FROM FILE CORFIL
C
        IF(IPRINT .EQ. 0) GO TO 160
```

```
C
      WRITE(IPRITE,6666)
 6666 FORMAT(1H1,////////,20X,"IPST CORE MATRIX -  NOT SCALED",//)
C
  160 CONTINUE
C
      DO 200 N5=1,LEVCOR(5)
          DO 210 N4=1,LEVCOR(4)
              DO 220 N3=1,LEVCOR(3)
                DO 230 N2=1,LEVCOR(2)
                      READ(IREAD2,33)(CORFST(N1,N2,N3,N4,N5)
     1                ,N1=1,LEVCOR(1))
    33                FORMAT(5F)
C
                      IF(IPRINT.EQ.0)GO TO 230
C
                      WRITE(IPITE,7777) (N1,N2,N3,N4,N5,
     1                CORFST(N1,N2,N3,N4,N5),
     2                N1=1,LEVCOR(1))
 7777                 FORMAT(1X,5(5I2,F15.0,1X))
  230               CONTINUE
  220           CONTINUE
  210       CONTINUE
  200   CONTINUE
C
C       SPECIAL OUTPUT
C
        CALL OUTPUT(1)
C
C       CLOSE FILE CORFIL
C
        CLOSE(UNIT=IREAD2)
C
        CALL MARGEN
C
C       OPEN INITIAL CORE MATRIX FILE CORINI
C
        OPEN(UNIT=IREAD3,DEVICE="SCIA",FILE=CORINI,ACCESS="SEQIN")
C
        CALL PRSCAL
C
C       CLOSE FILE CORINI
C
        CLOSE(UNIT=IREAD3)
C
        CALL CHISQR
C
C       CLOSE OUTPUT FILE
C
        CLOSE(UNIT=IRITE)
C
        STOP
        END
```

```
      SUBROUTINE MARGEN
C
C     MARGIN GENERATOR SUBROUTINE      STORED ON FILE MARGN.FOR
C     CREATED AUGUST 16, 1982          REVISED AUGUST 16, 1982
C     THIS SUBROUTINE USES A GIVEN CORE MATRIX TO CALCULATE DESIRED MARGINALS
C
C           DEFINITION OF VARIABLES
C                (TO BE FILLED IN)
C
      INTEGER CORRDM
C
      REAL MARG
C
      DIMENSION N(5),M(4)
      DIMENSION TMPMAR(10,5,5,5,5)
C
      COMMON /CORES1/ CORE(5,5,5,5,5)
      COMMON /MARGIN/ NDIM,LEVCOR(0/5),NMARG,MARDIM(10),CORRDM(10,5),
     1  MARG(10,5,5,5,5)
      COMMON/INOUT/ IREAD,IREAD1,IREAD2,IREAD3,ITYPE,IRITE,IRITE1,
     1  IRITE2
      COMMON /PRINT/ IPRINT,KPRINT
      COMMON/ISCALE/ SCALE
C
C
C  COMPUTE MARGINALS FROM CORE MATRIX
C
      DO 503 J=1,NMARG
C
      IF(MARDIM(J).EQ.4)GOTO 460
C
C     SET UNUSED DIMENSION INDICES OF MARGINALS TO 1
C
      DO 450 K=MARDIM(J)+1,4
          M(K)=1
  450 CONTINUE
C
C  SET POSSIBLY UNUSED DIMENSION OF CORE INDICES TO 1
C
  460 DO 470 K=1,5
          N(K)=1
  470 CONTINUE
C
C  LOOP THRU ALL CORE VALUES AND COMPUTE MARGINAL VALUES FROM CORE
C
      DO 480 N5=1,LEVCOR(5)
        N(5)=N5
        DO 490 N4=1,LEVCOR(4)
          N(4)=N4
          DO 500 N3=1,LEVCOR(3)
            N(3)=N3
            DO 510 N2=1,LEVCOR(2)
              N(2)=N2
              DO 520 N1=1,LEVCOR(1)
                N(1)=N1
                DO 530 K=1,MARDIM(J)
                  M(K)=N(CORRDM(J,K))
  530           CONTINUE
                MARG(J,M(1),M(2),M(3),M(4))
     1          =MARG(J,M(1),M(2),M(3),M(4))+
     2          CORE(N1,N2,N3,N4,N5)
```

```
  520                   CONTINUE
  510                CONTINUE
  500             CONTINUE
  490          CONTINUE
  480       CONTINUE
C
  600       CONTINUE
C
            WRITE(IRITE,8888) SCALE
 8888       FORMAT(1H1,//,20X,'MARGINALS - SCALED BY FACTOR ',E15.7,/////)
C
            DO 300 J=1,NMARG
                  DO 310 M4=1,LEVCOR(CORRDM(J,4))
                        DO 320 M3=1,LEVCOR(CORRDM(J,3))
                              DO 330 M2=1,LEVCOR(CORRDM(J,2))
                                    DO 340 M1=1,LEVCOR(CORRDM(J,1))
                                          TMPMAP(J,M1,M2,M3,M4)=
     1                                    MARG(J,M1,M2,M3,M4)
  340                               CONTINUE
                                    WRITE(IRITE,7777) (J,M1,M2,M3,M4,
     1                              TMPMAR(J,M1,M2,M3,M4),
     2                              M1=1,LEVCOR(CORRDM(J,1)))
 7777                               FORMAT(1X,5(5I2,F15.0,1X))
  330                         CONTINUE
  320                   CONTINUE
  310             CONTINUE
                  WRITE(IRITE,9999)
 9999             FORMAT(/)
  300       CONTINUE
C
            RETURN
            END
```

```
      SUBROUTINE PRSCAL
C
C     SUBROUTINE PRSCAL       STORED ON FILE PROSCL.FOR
C     SUBROUTINE TO DO ITERATIVE PROPORTIONAL SCALING.      (IPF METHOD)
C     READS FROM A FILE AN INITIAL CORE MATRIX WITH THE MAXIMUM NUMBER OF
C     DIMENSIONS 5 AND A MAXIMUM NUMBER OF 5 LEVELS PER DIMENSION.
C     MARGINAL MATRICES OF LESSER DIMENSIONS THAN THE CORE ARE
C     OBTAINED FROM SUBROUTINE MARGEN, AND THE CORE IS ITERATIVELY SCALED TO
C     THE MARGINALS.  THE PROGRAM DOES NOT CHECK THAT THE MARGINALS
C     ARE CONSISTENT.
C
C           DEFINITION OF VARIABLES
C                 (TO BE FILLED IN)
C
      INTEGER ENDFLG
C
      DIMENSION COROLD(5,5,5,5,5)
      DIMENSION CORTMP(5,5,5,5,5)
      DIMENSION COMARG(5,5,5,5)
      DIMENSION N(5),M(4)
C
      INTEGER CORRDM
C
      REAL MARG
C
      COMMON /CORES2/ CORE(5,5,5,5,5)
      COMMON /MARGIN/ NDIM,LEVCOR(0/5),NMARG,MARDIM(10),CORRDM(10,5),
     1 MARG(10,5,5,5,5)
      COMMON/INOUT/ IREAD,IREAD1,IREAD2,IREAD3,ITYPE,IRITE,IRITE1,
     1 IRITE2
      COMMON /PRINT/ IPRINT,KPRINT
      COMMON/ISCALE/ SCALE
C
      WRITE(ITYPE,999)
  999 FORMAT(1X,'TYPE IN 1 IF YOU WISH COMARG TO BE PRINTED',/
     1 ,1X,'OTHERWISE TYPE IN 0')
      READ(IREAD,11) KPRINT
   11 FORMAT(10I2)
C
C     READ IN INITIAL CORE MATRIX FROM FILE CORINI
C
      IF(IPRINT .EQ. 0) GO TO 60
C
      WRITE(IRITE,3333)
 3333 FORMAT(1H1,///,20X,'INITIAL CORE MATRIX - NOT SCALED',/////)
C
   60 CONTINUE
C
      DO 100 N5=1,LEVCOR(5)
             DO 110 N4=1,LEVCOR(4)
                    DO 120 N3=1,LEVCOR(3)
                           DO 130 N2=1,LEVCOR(2)
                                  READ(IREAD3,33)(CORE(N1,N2,N3,N4,N5)
     1                            ,N1=1,LEVCOR(1))
   33                             FORMAT(5F)
C
                                  IF(IPRINT.EQ.0)GO TO 130
C
                                  WRITE(IRITE,4444) (N1,N2,N3,N4,N5,
     1                            CORE(N1,N2,N3,N4,N5),
     2                            N1=1,LEVCOR(1))
```

```
 4444                                    FORMAT(1X,5(5I2,F15.0,1X))
  130                                CONTINUE
  120                            CONTINUE
  110                        CONTINUE
  100   CONTINUE
C
C   COMPUTE MARGINALS FROM CORE AND SCALE CORE
C   CONTINUE UNTIL CONVERGENCE OR MAXIMUM NUMBER OF ITERATIONS
C   HAS BEEN REACHED.  EACH ITERATION LOOPS THRU ALL MARGINALS
C
      WRITE(ITYPE,666)
  666 FORMAT(//,1X,'ENTER MAXIMUM NUMBER OF ITERATIONS',5X,$)
      READ(IREAD,66) MAXIT
   66 FORMAT(I5)
      WRITE(ITYPE,777)
  777 FORMAT(//,1X,'ENTER EPSIL',5X,$)
      READ(IREAD,77) EPSIL
   77 FORMAT(F10.3)
C
      WRITE(IRITE,55555) MAXIT,EPSIL
55555 FORMAT(1H1,///,' MAXIT=',I5,5X,'EPSIL=',E15.7,//)
C
      ITER=0
  400 ENDFLG=1
      ITER=ITER+1
C
      IF(ITER .EQ. 1) ENDFLG = 0
C
      DIFMAX = 0.
      N1DIF  = 0
      N2DIF  = 0
      N3DIF  = 0
      N4DIF  = 0
      N5DIF  = 0
C
      RATMAX = 0.
      N1RAT  = 0
      N2RAT  = 0
      N3RAT  = 0
      N4RAT  = 0
      N5RAT  = 0
C
      DELMAX = 0.
      N1DEL  = 0
      N2DEL  = 0
      N3DEL  = 0
      N4DEL  = 0
      N5DEL  = 0
C
      PRUMAX = 0.
      N1PRO  = 0
      N2PRO  = 0
      N3PRO  = 0
      N4PRU  = 0
      N5PRO  = 0
C
      DO 600 J=1,NMARG
C
C   REINITIALIZE COMPUTED MARGINAL MATRIX
C
      DO 410 I1=1,5
```

```
                DO 420 I2=1,5
                    DO 430 I3=1,5
                        DO 440 I4=1,5
                            COMARG(I1,I2,I3,I4)=0.
  440                   CONTINUE
  430               CONTINUE
  420           CONTINUE
  410       CONTINUE
C
            IF(MARDIM(J).EQ.4)GOTO 460
C
C           SET UNUSED DIMENSION INDICES OF MARGINALS TO 1
C
            DO 450 K=MARDIM(J)+1,4
              M(K)=1
  450       CONTINUE
C
C     SET POSSIBLY UNUSED DIMENSION OF CORE INDICES TO 1
C
  460       DO 470 K=3,5
              N(K)=1
  470       CONTINUE
C
C     LOOP THRU ALL CORE VALUES AND COMPUTE MARGINAL VALUES FROM CORE
C
            IF(KPRINT .EQ. 0) GO TO 475
C
            WRITE(IRITE,11111)
11111       FORMAT(1H1,////,20X,'COMARG',//)
C
  475       CONTINUE
C
            DO 480 N5=1,LEVCOR(5)
              N(5)=N5
              DO 490 N4=1,LEVCOR(4)
                N(4)=N4
                DO 500 N3=1,LEVCOR(3)
                  N(3)=N3
                  DO 510 N2=1,LEVCOR(2)
                    N(2)=N2
                    DO 520 N1=1,LEVCOR(1)
                      N(1)=N1
                      DO 530 K=1,MARDIM(J)
                        M(K)=N(CORRDM(J,K))
  530                 CONTINUE
                      COMARG(M(1),M(2),M(3),M(4))
     1                =COMARG(M(1),M(2),M(3),M(4))+
     2                CORE(N1,N2,N3,N4,N5)
                      IF(KPRINT.EQ.0) GO TO 520
                      WRITE(IRITE,44444)
     1                M(1),M(2),M(3),M(4),
     2                COMARG(M(1),M(2),M(3),M(4))
44444                 FORMAT(1X,5(5I2,E15.7,1X))
  520               CONTINUE
  510             CONTINUE
  500           CONTINUE
  490         CONTINUE
  480       CONTINUE
C
C     SCALE CORE BY GIVEN MARGINAL DIVIDED BY COMPUTED MARGINAL
C
```

```
      DO 540 N5=1,LEVCOR(5)
          N(5)=N5
          DO 550 N4=1,LEVCOR(4)
              N(4)=N4
              DO 560 N3=1,LEVCOR(3)
                  N(3)=N3
                  DO 570 N2=1,LEVCOR(2)
                      N(2)=N2
                      DO 580 N1=1,LEVCOR(1)
                          N(1)=N1
                          DO 590 K=1,MARDIM(J)
                              M(K)=N(CORRDM(J,K))
  590                     CONTINUE
                          SFACT=MARG(J,M(1),M(2),
     1                    M(3),M(4))/COMARG(
     2                    M(1),M(2),M(3),M(4))
                          CORVAL=CORE(N1,N2,N3,N4,N5)
                          CORNEW=CORVAL*SFACT
                          DIFF=ABS(CORNEW-CORVAL)
                          IF(DIFF.LE.DIFMAX)GO TO 592
                          DIFMAX=DIFF
                          N1DIF=N1
                          N2DIF=N2
                          N3DIF=N3
                          N4DIF=N4
                          N5DIF=N5
C
  592                     CONTINUE
                          RATIO=CORNEW/CORVAL
                          IF(RATIO.LE.RATMAX)GO TO 594
                          RATMAX=RATIO
                          N1RAT=N1
                          N2RAT=N2
                          N3RAT=N3
                          N4RAT=N4
                          N5RAT=N5
C
  594                     CONTINUE
                          IF(J.NE.NMARG)GO TO 599
C
                          IF(ITER.GT.1)GO TO 595
                          COROLD(N1,N2,N3,N4,N5)=CORNEW
                          GO TO 599
C
  595                     CONTINUE
                          DIFF=ABS(CORNEW-COROLD(N1,N2,N3,N4,N5))
                          IF(DIFF.LE.DELMAX)GO TO 596
                          DELMAX=DIFF
                          N1DEL=N1
                          N2DEL=N2
                          N3DEL=N3
                          N4DEL=N4
                          N5DEL=N5
C
  596                     CONTINUE
                          RATIO=CORNEW/COROLD(N1,N2,N3,N4,N5)
                          IF(RATIO.LE.PROMAX)GO TO 598
                          PROMAX=RATIO
                          N1PRO=N1
                          N2PRO=N2
                          N3PRO=N3
```

```
                        N4PRO=N4
                        N5PRO=N5
C
  598                   CONTINUE
                        IF(ABS(CORNEW-COROLD(N1,N2,N3,N4,N5))
     1                  .LE.EPSIL)GO TO 597
                        ENDFLG=0
C
  597                   CONTINUE
                        COROLD(N1,N2,N3,N4,N5)=CORNEW
C
  599                   CONTINUE
                        CORE(N1,N2,N3,N4,N5)=CORNEW
  580                CONTINUE
  570             CONTINUE
  560          CONTINUE
  550       CONTINUE
  540    CONTINUE
  600    CONTINUE
         IF(ENDFLG.EQ.1)GOTO 700          !HAVE FINAL CORE MATRIX
         IF(ITER.LT.MAXIT)GO TO 400       !MAXIMUM # OF ITERATIONS NOT REACHED
C
C   MAXIMUM NUMBER OF ITERATIONS REACHED OR METHOD CONVERGES
C
  700    CONTINUE
C
C   TYPE OUT FINAL CORE VALUES
C
         WRITE(IRITE,4411)
 4411    FORMAT(20X,'COMPARISON OF FINAL MARGINALS',
     1   ' FOR TWO CONSECUTIVE ITERATIONS')
         WRITE(IRITE,4422) ITER
 4422    FORMAT(/,5X,'NUMBER OF ITERATIONS=',I5)
         WRITE(IRITE,4433) DELMAX,N1DEL,N2DEL,N3DEL,N4DEL,N5DEL
 4433    FORMAT(/,5X,'ABSOLUTE VALUE OF MAX DEVIATION IS ',E15.7,
     1   ' AT POINT ',5I3)
         WRITE(IRITE,4455) PROMAX,N1PRO,N2PRO,N3PRO,N4PRO,N5PRO
 4455    FORMAT(/,5X,'MAX RATIO IS ',E15.7,' AT POINT ',5I3)
         WRITE(IRITE,4466)
 4466    FORMAT(///,20X,'COMPARISON OF TWO CONSECUTIVE MARGINALS')
         WRITE(IRITE,4433) DIFMAX,N1DIF,N2DIF,N3DIF,N4DIF,N5DIF
         WRITE(IRITE,4455) RATMAX,N1RAT,N2RAT,N3RAT,N4RAT,N5RAT
C
C        SPECIAL OUTPUT
C
         CALL OUTPUT(2)
C
         WRITE(IRITE,4477) SCALE
 4477    FORMAT(1H1,///,1X,'FINAL CORE MATRIX - SCALED BY FACTOR ',
     1   E15.7,////)
C
C        SCALE MATRIX FOR FINAL PRINTOUT
C
         DO 61) N5=1,LEVCOR(5)
            DO 620 N4=1,LEVCOR(4)
               DO 630 N3=1,LEVCOR(3)
                  DO 640 N2=1,LEVCOR(2)
                     DO 650 N1=1,LEVCOR(1)
                        COPTMP(N1,N2,N3,N4,N5)
     1                  =SCALE*COPE(N1,N2,N3,N4,N5)
  650                CONTINUE
```

```
  640              CONTINUE
  630            CONTINUE
  620          CONTINUE
  610    CONTINUE
C
C       PRINT OUT FINAL CORE MATRIX SCALED BY FACTOR SCALE
C
        DO 710 N5=1,LEVCOR(5)
               DO 720 N4=1,LEVCOR(4)
                      DO 730 N3=1,LEVCOR(3)
                             DO 740 N2=1,LEVCOR(2)
                                    WRITE(IRITE,4444) (N1,N2,N3,N4,N5,
     1                              CORTMP(N1,N2,N3,N4,N5),
     2                              N1=1,LEVCOR(1))
  740                        CONTINUE
  730                  CONTINUE
  720            CONTINUE
  710    CONTINUE
         RETURN
         END
```

```
      SUBROUTINE CHISQR
C
C     SUBROUTINE CHISQR       STORED ON FILE CHISQR.FOR
C
C     SUBROUTINE TO OBTAIN STATISTICS FOR ORIGINAL CORE VERSUS NEW CORE
C
      COMMON /CORES1/ COROLD(5,5,5,5,5)
      COMMON /CORES2/ CORNEW(5,5,5,5,5)
      COMMON /MARGIN/ NDIM,LEVCOR(0/5),NMARG,MARDIM(10),CORRDM(10,5),
     1 MARG(10,5,5,5,5)
      COMMON/INOUT/ IREAD,IREAD1,IREAD2,IREAD3,ITYPE,IRITE,IRITE1,
     1 IRITE2
C
C
      EPSIL = 1.0E-10
C
      SUM1 = 0.
      SUM2 = 0.
C
      DO 10 N5 = 1,LEVCOR(5)
          DO 20 N4=1,LEVCOR(4)
              DO 30 N3=1,LEVCOR(3)
                  DO 40 N2=1,LEVCOR(2)
                      DO 50 N1=1,LEVCOR(1)
                          X = AMAX1(COROLD(N1,N2,N3,N4,N5),EPSIL)
                          Y = AMAX1(CORNEW(N1,N2,N3,N4,N5),EPSIL)
C
                          D1 = ((X-Y)**2)/Y
                          D2 = X * ALOG(X/Y)
                          D3 = D2 + Y - X
C
                          SUM1 = SUM1 + D1
                          SUM2 = SUM2 + D2
   50                 CONTINUE
   40                 CONTINUE
   30             CONTINUE
   20         CONTINUE
   10     CONTINUE
C
      SUM2 = 2. * SUM2
C
      WRITE(IRITE,11) SUM1,SUM2
   11 FORMAT(1H1,'CHI SQUARE = ',E15.7,/,1X,'G SQUARE = ',E15.7)
      RETURN
      END
```

```
      SUBROUTINE OUTPUT(IS*)
C
C     DUMMY SUBROUTINE FOR SPECIAL OUTPUT
C
      RETURN
      END
```

References

1. Bishop, Y.M.M., S.E. Fienberg, P.W. Holland, "Discrete Multivariate Analysis," MIT Press, Cambridge, MA, 1975.

2. Goodman, L.A., J. Magidson (ed.), "Analyzing Qualitative/Categorical Data," Abt Books, Abt Associates, Inc., Cambridge, MA, 1978.

3. C.T. Ireland and S. Kullback, "Contingency Tables with Given Marginals", Biometrika (1968) 55, P. 179-188.

4. Dempster, A.P., N.M. Laird and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," J. Roy, Statist. Soc. Ser. B, 39, 1-38, 1977.

5. J.N. Darroch "Interactions in Multifactor Contingency Tables", J. Roy Statist Soc Ser B (1962) 24 p 251-263.

6. S.E. Fienberg "An Iterative Procedure for Estimation in Contingency Tables" The Annals of Mathematical Statistics (1970) 41, 907-917.

7. H.O. Hartley "Maximum Likelihood Estimation from Incomplete Data", Biometrics (1958) 14, 174-195.

8. Efron, Bradley, The Jackknife, the Bootstrap, and Other Resampling Plans, Monograph No. 38, Society for Industrial and Applied Mathematics, 1982.

U.S. Department
of Transportation

**Research and
Special Programs
Administration**

Kendall Square
Cambridge, Massachusetts 02142