Mining the Data Warehouse: Statistical Analysis of Combined Tables of Categorical Data

Gilbert W. Fellingham and H. Dennis Tolley

January 3, 1997

Abstract

We illustrate a maximum likelihood method for parameter estimation in combined tables of categorical data. Often, not all cells will contain data when multiple tables are combined. Also, since the same level of aggregation over covariates is often not available in each of the constituent tables some data will exist only on the margins. The method we present is appropriate when data for some cells are missing, and even when data may be available only on the margins. To illustrate the method we combine mortality tables from three different sources with different classification information. The code for the analysis is presented in SAS Proc IML, a powerful developmental tool for this type of analysis.

1 Introduction

The purpose of this paper is to demonstrate a systematic method of combining data from different tables. With the growing emphasis on storing, maintaining, and accessing data bases gathered over time and space, the need for statistical methodology to analyze such data is becoming increasingly critical. When gathered under the auspices of different managers over time and in different locations, even when the purpose underlying the data gathering is the same, seldom will the same or even similar protocols be utilized. Thus, data sets gathered for the same purpose may not (in fact, probably will not) be capable of being analyzed together using standard statistical programs.

The biggest problem with such data is level of data aggregation. For example, in the analysis which we show here, one data set has information on smoking status, one set includes only those who currently do not smoke, while one set has no information on smoking status. Although there is a growing collection of such data arising from many sources, combining such data has traditionally been done either using *ad hoc* methods or by fitting each data set to a model and then combining models. Assumptions implicit in accepting the resulting models are often hard to state. In the last decade, however, formal methods of combining data have been developed by theoretical statisticians. These methods can now be implemented to provide data analysts with additional tools for dealing with multiple sources of data.

We will illustrate a method of combining data from several tables into a single information table. The method illustrated uses maximum likelihood procedures to model the combined data. These likelihood procedures are based on the methodology presented in Tolley, Fellingham, and Scott (1996) who provide both estimation techniques and address identifiability problems when some of the data combined are aggregated at different levels for some tables relative to other tables. In this paper we show how to implement these procedures in practice.

2 A Concept Fixing Example

To fix ideas, consider the following three sets of data.

1. National. This data set was constructed as a synthetic data set, for illustration, using the Illustrative Life Table contained in Bowers, Gerber, Hickman, Jones, and Nesbitt (1986). These data are aggregated over gender, smoking status, and alcohol consumption status. Mortality data by age was available, and grouped into eleven five-year age classifications starting with ages 30-34 and ending with 80-84. There were 22 observations from this data set. This data set is shown in Table 1.

2. Church. This set consists of a ten year mortality experience study on insured individuals working for organizations associated with a church. The experience was between 1981 and 1991. As a

Table 1: The National data by status and age.

		Dead	Alive
Age	30–34	669	480417
	35–39	882	476519
	40–44	1327	470884
	45–49	2073	462157
	50–54	3178	448594
	55–59	4622	428537
	60–64	6619	399669
	65–69	8859	359914
	70–74	11449	307888
	75–79	13619	243967
	80–84	15220	170447

condition of employment, individuals could not be current users of alcohol or tobacco, although they might be former users. Both gender and age information were available. Although age was tabulated in single year age groups, for simplicity we divided age into the same five-year classifications as were used in the National data set. The data were predominantly for individuals in the U.S. though there were some employees residing in Canada. There were 44 observations from this data set.

3. CPSI. This set consists of the CPS-I data set from a study undertaken in the U.S. by the American Cancer Society. This data contains mortality outcomes for a group of people followed prospectively in time (Lew and Garfinkel, 1987). Individuals were enlisted into the study on a voluntary basis with no conditions of representativeness or randomness made. For each individual enlisted, smoking history was determined as well as age and gender. Age was tabulated in nine five-year age groups from 35-39 through 75-79. The study was initiated in 1959 and mortality data is available to 1972. There were 144 observations from this data set.

The problem considered in this paper is to combine these three mortality data files to form an estimate of mortality for each classification (i.e., age, gender, alcohol, and smoking status). We recognize that one of the data sets is artificial (Bowers et al. (1986)), and one is dated (Lew and Garfinkel (1987)). Therefore, the following example is more for illustration of the method than it is for determining real mortality patterns or related actuarial functions. Since data is available at different levels of aggregation as regards smoking status and alcohol consumption for the various groups, it is necessary to borrow information from some of the mortality studies to apply to the others. In this paper we illustrate how the likelihood solution of Tolley et al. (1996) can be implemented in combining life table data. This solution is based on the Poisson distribution and takes into account constraints imposed both by the sampling method and by the classification pattern of the data.

3 Preliminaries

We assume that each of the data sets to be combined can be put into the format of a general contingency table. The contingency table template will be the one defined by the classification variables available in any of the data sets. The levels of the classification variables will be those that are the least aggregated in the data sets. In the example, since information is available on smoking status, alcohol consumption status, gender, age and mortality outcome, we use a five-way contingency table. Note that not all classification variables are available for each data set. For example, data set one has no gender, smoking status, or alcohol consumption status. Data set two consists only of currently nondrinking, nonsmoking individuals, although these could be either never or former smokers, and never or former drinkers. In these cases, the information available is only the marginal information aggregated across some of the levels of the classification variables. We set up the table as if data were present in every cell. This not only helps the analyst understand the constraints inherent in the data, but aids in forming the components required in the methodology. In this case we have the following levels represented: (1) age group - 11 levels, (2) status - alive or dead, two levels, (3) gender - two levels, (4) smoking - never, former, current-moderate, current-heavy, four levels, (5) alcohol - no, yes, two levels, and (6) study - three levels. If data were present in every possible cell, we would have 1056 cells with data.

If we denote $\mathbf{c}$ as the vector of (hypothetical) data in each cell, then we assume

$$E[\mathbf{c}] = \mathbf{m}$$

where $\mathbf{c}$ is distributed as a Poisson random variable. We also assume that the vector $\mathbf{m}$ can be expressed as $e^{\mathbf{X}\boldsymbol{\beta}}$, the traditional log-linear model formulation Bishop, Fienberg, and Holland (1975). We now define a matrix $\mathbf{A}$ which we use to construct linear combinations of the logs of the cell counts. We can now write

$$\boldsymbol{\beta} = \mathbf{A}log(\mathbf{m}),$$

where $log(\mathbf{m})$ is the vector of natural logarithms of the individual components of $\mathbf{m}$. If $\mathbf{A}$ is full rank, then we can write,

$$log(\mathbf{m}) = \mathbf{X}\boldsymbol{\beta} = \mathbf{A}^{-1}\mathbf{A}log(\mathbf{m}).$$

In this case, if $\mathbf{X}$ is nonsingular then $\mathbf{A}^{-1}$ resembles the design matrix $\mathbf{X}$ and $\mathbf{A}log(\mathbf{m})$ represents the linear combinations of the logarithms of the expected counts that define the entries of $\boldsymbol{\beta}$. Often the matrix $\mathbf{X}$ will have full column rank but will not be square. As will be seen below, for combined data sets this is almost always true. In this case, the matrix $\mathbf{A}^{-1}$ has redundant columns that represent constraints which are implicit in $\mathbf{X}$. In other words, not all the $\boldsymbol{\beta}'s$ estimable in the complete data case will be estimable in the case we consider here. We will demonstrate below how to eliminate these degrees of freedom (terms in the model) to assure estimability of the likelihood. We refer the reader interested in more mathematical detail to the paper by Tolley et al. (1996).

4 The Analysis Matrices

We now present in detail certain matrix notation which is necessary for the analysis. First recall $\mathbf{m}$, representing the vector of expected cell counts, must have dimension corresponding to the possible number of cells. For our example, $\mathbf{m}$ will be of dimension 1056×1. Although this vector is hypothetical, it is very important to keep track of the subscripts associated with each element of $\mathbf{m}$. In our example, we used the following notation. Each element of $\mathbf{m}$ is designated $m_{i,j,k,l,m,n}$, where the subscripts are as follows: i, study number, $i = 1(National), 2(Church), 3(CPSI)$; j, alcohol use, $j = 1(no), 2(yes)$; k, tobacco use, $k = 1(never), 2(former), 3(moderate), 4(heavy)$; l, gender, $l = 1(male), 2(female)$; m, status (alive or dead), $m = 1(alive), 2(dead)$; n, age group, $n = 1(30-34), 2(35-39), \ldots, 11(80-84)$. We assume the subscripts move fastest at the far right (lexicographical order), so that the first 11 expected counts are for study 1, alcohol use 1, tobacco use 1, gender 1, status 1, and age groups $1, \ldots, 11$. The next 11 observations would be from status group 2, and so on. It is important to keep track of the relative positions of the various cells in the $\mathbf{m}$ vector so that other necessary matrices which represent linear combinations of $\mathbf{m}$ will be constructed appropriately.

The actual data is kept in a vector which we call $\mathbf{n}$. For these three data sets, the entire data vector $\mathbf{n}$ has dimension 210×1. That is, there are 210 observations available for the analysis. The vector of observations for this analysis was formed by stacking the columns of data in Tables 1, 2, and 3. So the first element of the vector $\mathbf{n}$ is 669, the second is 882, the twelfth is 480417, etc.

We define a matrix $\mathbf{W}$ to identify the cells from which the actual observations were drawn. The $\mathbf{W}$ matrix has a row corresponding to each or the 210 observed counts and a column corresponding to each of the 1056 cells in the complete combined contingency table. Thus, the dimensions of the $\mathbf{W}$ matrix for this problem are 210×1056. In this case, none of the studies supplied data for individual cells, all data were from margins. Building the $\mathbf{W}$ matrix for this particular problem is mainly a bookkeeping issue. Each row of the $\mathbf{W}$ matrix consists of zeros and ones, with a one corresponding to each cell which contributed to the total count in the observation represented by the row. In the complete data case, the $\mathbf{W}$ matrix would be an identity, representing one observation from each cell. In this problem, each row contained multiple ones, indicating each observation represented a sum over a number of different cells. Thus, $\mathbf{Wm}$ yields the expected counts of the actual data.

We now construct the first row of the $\mathbf{W}$ matrix explicitly. The first row of the $\mathbf{W}$ matrix must correspond to the first data element of the vector $\mathbf{n}$. This element is the 669 count of subjects in the National study whose status is *dead* and whose age is *30-34*. These subjects comprised both genders from all alcohol use and tobacco use levels. We now recall how the $\mathbf{m}$ vector was constructed. The first one-third (352 of the 1056 cells) of the elements of $\mathbf{m}$ represent the National study. Elements $1, 12, 23, \ldots$ represent ages $30-34$, elements $2, 13, 24, \ldots$ represent ages $35-39$, etc. Elements $1-11, 23-33, 45-55, \ldots$ represent *alive* while elements $12-22, 34-44, 56-66, \ldots$ represent *dead*. Gender would be represented in blocks of 22, with the first 22 cells representing males, the next 22 females, the next 22 males, etc. Tobacco use is in blocks of 44, with the first 44 elements for *never*, the next 44 for *former*, the next 44 for *moderate*, and the next 44 for *heavy*. That cycle repeats six times. Finally, non-alcohol users would be represented in the first 176 cells, alcohol users in the next 176 cells, with this cycle repeated three times. With this in mind the first observation of 669 would include people from cell 12 (*age* $30-34$,*status* dead, *gender* male, *tobacco* never, *alcohol* no, and *study* 1), cell 34 (*age* $30-34$,*status* dead, *gender* female, *tobacco* never, *alcohol* no, and *study* 1), cell 56 (*age* $30-34$,*status* dead, *gender* male, *tobacco* moderate, *alcohol* no, and *study* 1), etc. So the first row of the **W** matrix would consist of $1's$ in cells $12, 34, 56, \ldots, 342$, and $0's$ everywhere else. The next 209 rows of the **W** matrix are produced analogously. The SAS Proc IML code used to produce part of the **W** matrix is shown in Figure 1.

Figure 1: Code used to produce the **W** matrix.

```
/********************************
SAS code to produce the W matrix
********************************/
proc iml;

w=repeat(0,210,1056);

/**  first study (National)
     first 22 rows of w matrix
                                **/
do j=1 to 11;
  do i=11+j to 330+11+j by 22;
    w[j,i]=1;
  end;
end;

do j=12 to 22;
  do i=j-11 to 330+j-11 by 22;
    w[j,i]=1;
  end;
end;
```

To construct the **A** matrix, recall that this matrix is used to form linear combinations of the logs of the expected cell counts ($\mathbf{m}$). Thus, we must keep track of the elements of $\mathbf{m}$ as we construct **A** the same way that we did to construct **W**. These linear combinations will, hopefully, be those of most interest to the researcher. However, we need to keep in mind that because of data sparseness, not all linear combinations will be estimable.

Initially we work as if data were present in all cells. We require **A** to be nonsingular. We constructed our **A** matrix as follows. The first row computed the overall average of the log cell counts. This is simply a row of $\frac{1}{1056}'s$ which yields the average of the log cell counts when multiplied by $log(\mathbf{m})$. The next ten rows we used were orthogonal polynomials constructed to estimate linear, quadratic, cubic, etc., up to the 10^{th} degree functions of age. Orthogonal polynomials may be automatically constructed using the function `orpol` in SAS Proc IML SAS Institute (1990). Every 12^{th} cell the coefficients will start to repeat since there are 11 age groups. These coefficients will be repeated in 96 blocks of 11 along the rows of the **A** matrix since age moves the fastest in the $\mathbf{m}$ vector. The next row, row 12, computes the difference between the number of alive and the number of dead, or what we called the *status* effect. Since the $\mathbf{m}$ vector has 11 cells from *status*=alive followed by 11 cells from *status*=dead, the row of the **A** matrix will have 48 groups of 11 $-1's$ followed by 11 $1's$.

Row 13 estimates the gender effect. Since there are 22 cells associated with males followed by 22 cells associated with females in the $\mathbf{m}$ vector, this row of the **A** matrix will be 24 groups of 22 $-1's$ followed by 22 $1's$.

The next three rows (rows 14 through 16) were constructed to estimate the effect of smoking. The first contrasted level one (never smoked) with level two (former smoker) and was constructed as 6 groups of 44 $-1's$ followed by 44 $1's$ followed by 88 $0's$. The next row contrasted level three (moderate smoker) with level four (heavy smoker), and also had 6 groups, this time with 88 $0's$ followed by 44 $-1'$ followed by 44 $1's$. The final degree of freedom for smoking contrasted levels one and two (never and former smokers) with levels three and four (moderate and heavy smokers). This row consisted of 6 groups of 88 $-1's$ followed by 88 $1's$

The next row estimated the alcohol effect. This row consisted of 3 groups of 176 $-1's$ followed by 176 $1's$. The next two degrees of freedom were associated with the effect of *study*. Since the three data sets were gathered a number of years apart, these linear combinations were constructed to look for a time effect. The first of these contrasted the Church study (the most recent) against the CPSI study (the earliest). This row of the **A** matrix consisted of 352 $0's$ followed by 352 $1's$ followed by 352 $-1's$. The second degree of freedom for study was a quadratic effect of time, the Church and CPSI data contrasted against the National data (collected in between the other two). This row consisted of 352 $-2's$ followed by 704 $1's$. These were all the degrees of freedom for the main effects. The other degrees of freedom were constructed as interactions among these main effects.

The interaction rows of the **A** matrix are calculated from the main effects (and from previous interaction terms) as component wise products of the entries of these previous rows. This multiplication is accomplished quite easily using the `hdir` command in SAS Proc IML. Three-way and higher interaction terms are produced analogously. The dimensions of the **A** matrix for our example are 1056×1056.

The final matrix which needs to be understood is the **Z** matrix. This matrix describes which of the observed totals are constrained by sampling considerations. In these studies, the total alive plus dead is fixed, so these totals are constrained. The **Z** matrix is again mostly zeros, with ones in each row identifying which cell counts must total to a fixed number. Since the total over status was fixed, a row of the **Z** matrix will have a one in the column associated with "alive" and a one in the column associated with "dead" for each given pair of observations.

The **Z** matrix only operates on observed data, so as we build the **Z** matrix, we need only concern ourselves with the form of the observed data vector **n**. The first number in **n** is 669 and the twelfth number is 480417. Since these numbers represent the total number of subjects in the $30 - 34$ age group for the National study, their sum is a constrained total. Thus, the first row of the **Z** matrix will contain a 1 in the first and twelfth positions, and $0's$ elsewhere. There are 210 observed data points with 105 data pairs, so the dimensions of **Z** are 105×210.

5 Building the Model

Since the only data available are the marginal counts represented by **n**, and since we are subject to those constraints made explicit in the **Z** matrix, we are restricted as to which $\beta's$ are actually estimable. In fact, the likelihood function itself (see below) ultimately determines which of the $\beta's$ may actually be included in the final model.

In this analysis, since alive or dead status is a constrained total, we really have a bivariate measure in each of 528 cells. That is, if we think of the cells as the proportion of alive and dead in each setting, we are constrained that those proportions add up to 1. We are limited by that constraint to focus on the degree of freedom associated with the main effect for status, and all interactions involving status. These account for the 528 degrees of freedom which would be estimable if all data were present.

For the sake of simplicity, we also limited our search for estimable degrees of freedom to three-way interactions or lower and only to polynomials of up to the fifth degree involving age. This left our possible search space at 65 degrees of freedom. Since these are the only degrees of freedom we wish to examine, we define the matrix **B** to be the 65 columns of $\mathbf{A}^{-1}$ corresponding to these degrees of freedom. The **B** matrix is only of dimension 1056×65 and represents the design matrix **X**.

However, not even all of these 65 degrees of freedom will be estimable. A final determination of estimability is made by appealing to the Implicit Function Theorem (Apostol, 1957). As shown in Tolley et al. (1996) this results in assessing the rank of a matrix whose individual components are formed from the derivatives of the likelihood. Since we use Newton-Raphson as our estimation procedure, this matrix must be computed. Thus, if any of the degrees of freedom (terms of the model) are not estimable, the Jacobian will not be invertable and the computation will fail.

Since we use the Newton-Raphson method to estimate the $\beta's$, it is important to get initial estimates within the radius of convergence. This is not always a trivial matter. We are currently exploring numerical options which may make this step easier to implement. We built our models by first finding estimates for a small model which had essentially main effects (the two-way interactions involving status are main effects in this model), and then, using these estimates for the starting point, testing gradually more complicated models. We used $((\mathbf{WB})'\mathbf{WB})^{-1}(\mathbf{WB})'log(\mathbf{n})$ for the initial estimates, where **B** represents the columns taken from the $\mathbf{A}^{-1}$ matrix which were limited to those corresponding to the effects for *status, status by alcohol, status by tobacco, status by gender, and status by age(linear).* Figure 4 shows the section of SAS Proc IML code which was used for the actual implemenation of the Newton-Raphson method.

6 Results

Although we had some difficulty extracting the alcohol effect from the Church effect because only the Church data had alcohol information, this fit seemed reasonable. We show two plots of expected (based on the model) rates and actual rates based on the CPSI and National data. These plots were built using SAS/GRAPH.

7 Conclusion

We have illustrated a maximum likelihood technique which allows estimation of effects in log-linear models of categorical data even when many cells are missing data and some data may be available only on the margins. We have demonstrated the methodology using three data sets which have only marginal information, so none of the available data represents a count in a single cell. SAS Proc IML offers a powerful platform to perform this analysis.

Figure 2: The portion of SAS Proc IML code used to actually implement the Newton-Raphson algorithm. $\mathbf{W}$ is `w`, $\mathbf{Z}$ is `z`, `x` contains the columns of $\mathbf{A}^{-1}$ which are estimable ($\mathbf{B}$), and `b` is a vector of the initial estimates for the $\beta's$. The first derivative of the log-likelihood is designated `f`, and the second derivative is designated `j`.

```
g=nrow(z);
k=nrow(w);
con=100;
do while(con>.00000000001);

m=exp(x*b);

f=(n/(w*m))‘*(w*(x#m))
  -j(1,k,1)*(w*(x#m))
  -((z*n)/(z*(w*m)))‘*(z*(w*(m#x)))
  +j(1,g,1)*(z*(w*(m#x)));

j=-(w*(x#m))‘*((n/(w*m)##2)#(w*(x#m)))
  +((w*x)‘*(n/(w*m)#(w*(x#m))))
  -((w*x)‘*(w*(x#m)))
  +(z*(w*(x#m)))‘*(((z*n)/((z*(w*m))##2))
                    #(z*(w*(x#m))))
  -(((z‘*((z*n)/(z*(w*m))))#(w*x))‘
                    *(w*(x#m)))
  +(((z‘*j(g,1))#(w*x))‘*(w*(x#m)));

b1=b-inv(j)*f‘;
con=max(abs(b1-b));
b=b1;
end;
```

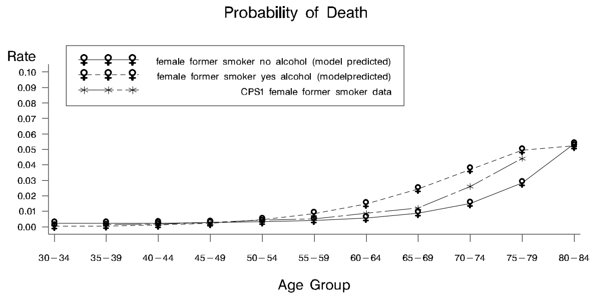

Figure 3: Expected mortality experience of previous smoking females and actual CPSI rates.

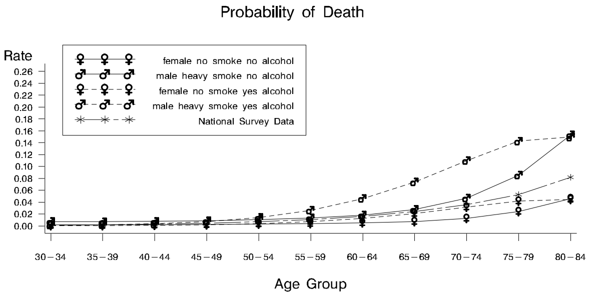

Figure 4: Expected mortality experience of never smoking females, heavy smoking males, and actual National rates.

References

Apostol, T. (1957). *Mathematical Analysis.* Reading, MA: Addison-Wesley Publishing Co., Inc.

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, MA: The MIT Press.

Bowers, N. L., Gerber, H. U., Hickman, J. C., Jones, D. A., and Nesbitt, C. J. (1986). *Actuarial Mathematics.* Itasca, IL: Society of Actuaries.

Lew, E. A. and Garfinkel, L. (1987). Differences in mortality and longevity by sex, smoking habits and health status with discussion. *Transactions of the Society of Actuaries*, 107–130.

SAS Institute, Inc. (1990). *SAS/IML Software: Usage and Reference, Version 6* (First edition). SAS Institute Inc.

Tolley, H. D., Fellingham, G. W., and Scott, D. T. (1996). Likelihood methods for combining tables of data. *Scandanavian Actuarial Journal* **Submitted**.