The Use of the SAS® Programming Language and Procedures to Model and Process Data Used in Creating the New U.S. Mortality Atlas

Gretchen K. Jones, Linda W. Pickle, and Michael Mungiole, CDC/NCHS, Hyattsville, Maryland

ABSTRACT:

Thematic maps have helped epidemiologists to identify cancer “hot spots” in the U.S. by providing a visualization of the geographic patterns of mortality not apparent from tabular statistics (Mason 1975, 1976; Pickle 1987, 1990). The National Center for Health Statistics has prepared an atlas of maps of death rates for the leading causes of death in the U.S. for the period 1988-1992. This project included fitting complex random effects models to examine patterns in mortality data for 80 race-sex/cause-specific datasets, each of which consisted of 8000 observations. We used the recently-released PROC MIXED to fit these data. In order to minimize programming changes and potential errors, we utilized SAS macros, concatenation of files, and other programming tools. In this paper we discuss the techniques used to repeatedly apply this complex model, producing self-identifying output files for subsequent processing.

INTRODUCTION

Maps of mortality rates (Mason 1975, 1976; Pickle 1987, 1990) have proven to be quite helpful in visualizing trends. Such maps have helped to identify “hot spots” and spatial patterns in data for the U.S. that are not apparent in tabulated data. Because of the usefulness of previously published atlases, the National Center for Health Statistics is publishing a mortality atlas on 18 major causes of death in the U.S.

For each cause and race-sex combination, four maps will be shown: a map of observed age-adjusted death rates; a map showing the significance of each rate compared to the U.S. rate; and two smoothed age-specific maps of predicted rates. Together these maps will allow the reader to read off a rate for an HSA, to determine the significance of each rate, and to observe general spatial patterns.

The mortality rates were age-adjusted, using the 1940 proportions of the population in each of 10 age groups in the United States. A direct age-adjustment method (Pickle & White 1995) was used because it allows valid comparison of mortality rates across areas with different age distributions without requiring assumptions of the independence of age and area effects. The geographic unit that was chosen for mapping was the Health Service Area (HSA) (Makuc 1991). This area represents an aggregation of counties based on where the residents obtain their hospitalization.

In previously-published cancer atlases, the simple binomial variance has been used to test whether an area’s rate was significantly different from the U.S. rate (Chiang 1961; Mason 1975, 1976; Pickle 1987, 1990). However, this variance estimator has been shown to often underestimate (Brillinger 1986), and occasionally overestimate (Pickle 1992) the true variance of the rate. Empirical Bayes methods have been proposed to “stabilize” rates with large variance (Manton 1989; Clayton and Caldor 1987), but these methods tend to over-shrink the individual rates to some overall rate, perhaps masking interesting spatial patterns. We have explored the use of random effects models to account for over-and under-dispersion in the data and to provide confidence limits for regional rates. Also, this model provides a better estimate of the sources of variance than the simpler binomial variance.

STEPS IN ATLAS PROCESSING

Figure 1 gives a conceptual flow of the atlas processing. First the rates, variances and counts were calculated from the basic Vital Statistics
Mortality Files. The statistics were accumulated for 19 causes, 805 HSAs and four race-sex groups. This was done on the mainframe, using SAS Data Step programming. The files were downloaded, rates and variances were reformatted using S+ (Statistical Sciences, Inc. 1992) and were input into Atlas Pro (Strategic Mapping Inc. 1992), which produced three types of rate maps: (1) the observed rates, (2) results of significance tests comparing rates for each HSA to the U.S. rate, and (3) smoothed rates.

A random effects model of mortality data was used to determine valid estimates of the rates. Spline models of the age-specific data were flexible enough to fit each of the cause-specific datasets well. In order to determine the form of the model to be used, we analyzed the U.S. data and calculated the log-linear and cubic spline best fits to the age-specific data. Once we determined the model form for each cause and race-sex group, we were ready to input the detailed data into the newly-released SAS PROC MIXED, to describe the model (in the MODEL statement), and to output files containing predicted HSA rates and predicted regional rates. SAS was also used to create marginal regional rates. Initial analysis of the death rates using PROC MIXED included weights equal to the number of deaths. However, when many of the observed numbers of deaths were small, results from the model were unstable. To overcome this, when the number of deaths was less than three, a revised weight was calculated based on the marginal rates in that region.

The next steps involved a smoothing algorithm called "headbanging," written in S+ (Hansen 1991). The smoothed predicted rates were then input into an Atlas Pro program, which produced the age-specific maps. Figure 2 is an example of one such map, lung cancer rates for middle-aged white males.

SAS CONSIDERATIONS

One of the tasks which faced us when we were testing the models on the cause-specific data (inputting the data into PROC MIXED, using different MODEL statements), was that of being able to submit different runs for 76 different permutations of race-sex and cause, all of which were to be processed the same except for the cause and race-sex differences. We wanted the DSname of the output file to contain identifying codes for cause and race-sex. We chose a four character mnemonic such as BF07, for Black Female, cause 07. We wanted to have a "master" program, or template, to which we could make as few changes as possible, and then submit to completely process the data for that cause and race-sex. With the use of macros, we were able, using a TSO global change of those four characters in the job (JCL plus the SAS program code), to completely adapt the job for a new cause-race-sex group.

EXAMPLE: The following coding will handle the execution of PROC MIXED for any cause-race-sex combination. The "parm" is set to BF07 in this example. Only the parm has to be changed to run a new cause-race-sex combination. This example also illustrates the use of CALL SYMPUT to create a macro-variable to represent the independent variables of each model according to a table of values covering all cause-race-sex combinations.

```
IF N = 1 THEN DO;
    PARM = 'BF07';
    KAUSE = SUBSTR(PARM,3,2);
    KRACESEX = SUBSTR(PARM,1,2);
    KDEC1=2;
    KDEC2=3;
    KDEC3=5;
    KDEC9=1;
    CDEC = KDEC(KAUSE);
    DO I=2 TO 9;
        IF I LE CDEC THEN DK(I) = 0;
        ELSE DK(I) = (I - CDEC)**3;
    END;
    CALL SYMPUT (MDK2',TRIM(LEFT(DK2)));
    CALL SYMPUT (MDK3',TRIM(LEFT(DK3)));
```
CALL SYMPUT
   ('MDK9',TRIM(LEFT(DK9)));
END;

PROC MIXED; BY REGION;
CLASSES HSA REGION;
MODEL LNRATE = DECADE DEC2 DEC3
   DECKNOT / PREDICTED S;
RANDOM INTERCEPT / SUBJECT =
   HSA (REGION);
WEIGHT WTUS;
ESTIMATE '0' INTERCEPT 1 DECADE .25
   DEC2 1 DEC3 1 DECKNOT 0;
ESTIMATE '1' INTERCEPT 1 DECADE 1
   DEC2 4 DEC3 8 DECKNOT &MDK2;
ESTIMATE '2' INTERCEPT 1 DECADE 2
   DEC2 9 DEC3 27 DECKNOT &MDK3;
ESTIMATE '9' INTERCEPT 1 DECADE 9
   DEC2 81 DEC3 729 DECKNOT &MDK9;

Another problem which we were able to solve
using SAS coding techniques was to include
needed identification variables (cause and race-
sex) in the “estimate” file, which did not
automatically appear in the output dataset created
by the “MAKE ‘ESTIMATE’ OUT=....”
statement. We first created a SAS data set
containing those ID variables only, and then we
concatenated that data set with the “estimate”
output dataset in the SET statement, read in the
ID file first, and retained those variables for the
subsequent reading of the estimate file.

DATA OUTDATA (KEEP = ........)
   SAVEIDS (KEEP = CAUSE RACESEX);
SET;
IF _N_ = 1 THEN OUTPUT SAVEIDS;

PROC MIXED; BY REGION;

MAKE 'ESTIMATE';
   OUT=OUT2.REGNOUT;

DATA OUT3.REGNBF07;
SET SAVEIDS REGNOUT;
RETAIN CAUSE RACESEX;
IF _N_ = 1 THEN RETURN;
OUTPUT;

SUMMARY

The authors, affiliated with CDC/NCHS (Centers
for Disease Control/National Center for Health
Statistics), were able to incorporate SAS
procedures and techniques into a very complex
processing procedure using basic Vital Statistics
Mortality Data from NCHS in order to produce
Figure 1.

Conceptual Flow Chart of Atlas Processing

Create basic rate maps
- DVS detail mort. tapes 1988-1993
- calc. rates, vars & counts (mainframe)
- Ratevar (to PC file) 1 obs/cause, racesex, HSA
- Download, sort by cause, HSA split by sex/race

Determine best fit (model form & spatial covariance) by analyzing U.S. data
- 1990 HSA population centroids
- U.S. totals summed from Ratvarwm.srt
- plot correlograms (S+)
- model corr's using TableCurve
- create file of rates & vars only (805x51) + field names (S+)
- Ratvarwm.csv 1 obs/HSA
- Atlas Script basic maps

Create significance maps
- calc. Poisson parameter MLEs, dispersion factors
- Ratmlewm.csv
- Atlas Script significance maps

Model rates
- Ratevar (mainframe) detailed data
- Hsavn.dat
- SAS MIXED random effects modeling (Mainframe)
- predicted HSA rates PREDWM01
- Predicted Regional rates RPGNWM01
- Calculate regional 95% CIs
- headbanging smoothing algorithm (S+)
- Analyze residuals (S+)
- smoothed predicted rates
- Plots & correlograms
- Plots & rowplots
- smoothed age-specific maps
- Atlas Script
REFERENCES


1078


SAS is a registered trademark or trademark of SAS Institute Inc. in the USA and other countries. © indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.